DANIEL L. STUFFLEBEAM
CHRIS L. S. CORYN

SECOND EDITION

Evaluation Theory, Models, & Applications

JOSSEY-BASS™
A Wiley Brand

EVALUATION THEORY, MODELS, AND APPLICATIONS

EVALUATION THEORY, MODELS, AND APPLICATIONS

Second Edition

Daniel L. Stufflebeam
Chris L. S. Coryn

JOSSEY-BASS™
A Wiley Brand

Cover design: Wiley
Cover image: © PASIEKA | Getty

Published by Jossey-Bass
A Wiley Brand
One Montgomery Street, Suite 1200, San Francisco, CA 94104-4594—www.josseybass.com

Library of Congress Cataloging-in-Publication Data

Stufflebeam, Daniel L.
Evaluation theory, models, and applications / Daniel L. Stufflebeam, Chris L.S. Coryn.—Second edition.
pages cm.—(Research methods for the social sciences ; 50)
Includes bibliographical references and index.
ISBN 978-1-118-07405-3 (cloth)—ISBN 978-1-118-87032-7 (pdf)—ISBN 978-1-118-87022-8 (epub)
1. Evaluation research (Social action programs) I. Coryn, Chris L. S. II. Title.
H62.S79627 2014
001.4—dc23

2014001638

Printed in the United States of America

SECOND EDITION

HB Printing 10 9 8 7 6 5 4 3 2 1

CONTENTS

LIST OF FIGURES, TABLES, AND EXHIBITS

Figures

Tables

Exhibits

Daniel L. Stufflebeam dedicates this book to Carolyn, his wife of fifty-seven years; Egon Guba, his revered, departed colleague; Dr. Anthony Shinkfield, his coauthor of previous books; and his many outstanding graduate students.

Chris L. S. Coryn dedicates this book to his wife, Dr. Daniela C. Schröter, and their daughter, Lilly Thea Coryn-Schröter; his numerous students; and Drs. Michael Scriven, E. Jane Davidson, and E. Brooks Applegate, his longtime mentors.

PREFACE

This second edition of *Evaluation Theory, Models, and Applications* builds on the original volume (Stufflebeam & Shinkfield, 2007) by presenting its core material and infusing new content from recent developments in evaluation. The book is grounded in my long involvement in helping develop the evaluation profession and also reflects the contributions of my new coauthor, Dr. Chris L. S. Coryn. It is intended for use as a textbook for graduate-level courses in program evaluation and as a tool for evaluators and administrators and other clients and users of evaluation.

In developing this edition, Dr. Coryn and I received and addressed constructive feedback from reviewers—especially those commissioned by Jossey-Bass—of both the original volume and this second edition's first draft. The reviewers asked for information on the perspective from which this edition was prepared, including its boundaries as well as its reach. They also stressed that this book would provide a valuable service by drawing out and sharing lessons from our rich array of involvements in evaluation.

Dr. Coryn and I do not claim, for this volume, all-encompassing scope and equitable balance in covering the full reach of evaluation's approaches, developments in all countries, and applications in the full range of disciplines. Actually, no one could do that for any textbook. Instead, we have written about what we have experienced as important and useful in evaluation work, based on our many and varied evaluation involvements and on those of leaders in the evaluation field. We hope readers will find that this book provides one significant and useful set of insights into the complex sphere of program evaluation.

We have designed this volume to offer a broad, experience-based perspective on the evaluation field's background, its theories and standards, its alternative approaches to evaluation, the extensive assortment of qualitative and quantitative procedures and management steps required to carry out sound evaluations, metaevaluations, and processes for institutionalizing and mainstreaming systematic evaluation. Given its breadth and depth of coverage, we see the book as appropriate for supporting at least two graduate-level courses on evaluation theory and practice.

The book's discussed historical period is approximately 1930 to 2014. Within that time frame, my (first author) perspective reflects my establishment of the internationally known Evaluation Center, directing it at The Ohio State University from 1963 to 1973, and continuing to direct it at Western Michigan University from 1973 to 2002. Based on a variety of evaluation experiences, I developed the CIPP model, which lays out a comprehensive approach to assessing a program's context, inputs, processes, and products. Although the CIPP model is the major approach advocated and explained throughout this book, several other leading

approaches are reviewed and advocated, including Scriven's consumer-oriented approach; Stake's responsive evaluation; Guba's constructivist, naturalistic evaluation; Patton's utilization-focused evaluation; experimental design; case study evaluation; and Tyler's objectives-based evaluation.

I helped develop what is now the American Evaluation Association. I also founded and directed the national Joint Committee on Standards for Educational Evaluation, which has produced North American professional standards for evaluations of programs, personnel, and students. Those standards stress that evaluations must be not only accurate but also useful, feasible, ethical, and accountable. The Joint Committee's standards are employed extensively in this book to help define what is meant by excellent evaluation practice and to provide foundational criteria for conducting metaevaluations.

As is patently clear throughout the book, many collaborations have beneficially shaped my ideas about evaluation. I have conducted evaluation work with many of evaluation's icons, especially Egon Guba, Michael Scriven, George Madaus, Ralph W. Tyler, Robert Stake, Jason Millman, Thomas Kellaghan, Sydney Pressey, Richard Jaeger, James Sanders, William Webster, Arlen Gullickson, John Hattie, Elliott Eisner, Jerry Horn, Bill Gephart, Ernest House, and Michael Coplen. My perspective has also been shaped by writing with many coauthors, especially Dr. Coryn (for this book) and Anthony Shinkfield (for the first edition). In addition, my views of evaluation have benefitted from collaborations with outstanding graduate students, including Howard Merriman, Thomas Owens, David Nevo, Blaine Worthen, Diane Reinhard, Robert Rodosky, Sandra Ryan, Sharon Dodson, P. Cristian Gugiu, Darrell Root, Jerry Walker, Gary Wegenke, Lori Wingate, Daniela C. Schröter, Jeri Ridings, and many more.

In contrast to the original book, this second edition beneficially reflects Dr. Coryn's evaluation background. He directs the world's only interdisciplinary PhD program in evaluation. Its students come from countries throughout the world to earn evaluation doctoral degrees based in such disciplines as sociology, social work, education, computer science, nursing, business, and engineering. He has conducted highly influential research on national government research offices throughout the world. And he edits the Web-based *Journal of MultiDisciplinary Evaluation*. I am exceedingly pleased with Dr. Coryn's contributions to this book and expect that readers will also appreciate the many up-to-date references to new evaluation material and the significant experiences that he has brought to the volume.

On the one hand, readers will note that many of the book's referenced examples are from education in the United States. This feature tends to narrow the book's perspective, especially for readers outside North America. On the other hand, Dr. Coryn and I have drawn on our experiences to cite examples from such fields as community and economic development, housing, community-based programs for youth, environmental protection programs, military personnel evaluation, and government programs to promote safety in the railroad industry. The book also draws lessons from evaluation experiences in such countries as Columbia, Ecuador, Scotland, Poland, Spain, Finland, Ireland, India, Malaysia, Thailand, the Philippines, Jamaica, Russia, Switzerland, Israel, New Zealand, and Australia. In the main, though, the book is heavily grounded in U.S. culture.

In developing this volume, Dr. Coryn and I have aimed to deliver a unique and significant offering among the growing number of evaluation textbooks. Clearly, this book is not a comprehensive, encyclopedic presentation of all important evaluation developments that have occurred everywhere in the world and across all disciplines and service areas. No evaluation textbook has met or could meet such a demanding set of requirements. Nevertheless, readers will find that this book's coverage is rich, deep, and extensive. I agree with our reviewers that the lessons Dr. Coryn and I have learned through an extensive array (and, in my case, over fifty years) of evaluation experiences are worthy of serious consideration by evaluation students, evaluation practitioners, and evaluation users. I hope evaluators, evaluation professors and students, evaluation sponsors, evaluation clients, administrators, and other evaluation stakeholders will find this book to be enlightening and useful for conducting and applying evaluations that help meet the needs associated with program improvement and accountability.

Daniel L. Stufflebeam

ACKNOWLEDGMENTS

The writing of this book has required the help and encouragement of many others. We thank Jossey-Bass, our publisher, for delineating so clearly, following its marketing analysis, what content should be emphasized in a book on program evaluation. In particular, we owe a special thanks to Jossey-Bass/Wiley senior editor Andrew Pasternack and associate editor Seth Schwartz for their ongoing support and encouragement and to Francie Jones for her expert assistance in editing the draft manuscript. We regret deeply Mr. Pasternack's untimely death.

We also thank the numerous doctoral students in the Interdisciplinary PhD in Evaluation (IDPE) program at Western Michigan University (WMU) and students from other WMU programs who field-trialed and provided valuable feedback on draft chapters. Special acknowledgment is given to Kristin Hobson for her meticulous checking, rechecking, and editing of the book's references, her thorough review of draft chapters, and her assistance in developing PowerPoint summaries and test items for chapters, and to Carl Westine for his invaluable input on early drafts of certain chapter sections. We have benefited greatly from the contributions of Kristin and Carl, who have shared their graduate student perspective. We are also mindful and appreciative of the substantive contributions by Anthony Shinkfield, who coauthored the original volume of this book. Clearly, many of his substantive contributions to that book have carried over into this one. Finally, we thank Lois-Ellin Datta, Randy Davies, and Jean King for their insightful, detailed, useful critiques of all chapters and the book as a whole.

Chris L. S. Coryn
Daniel L. Stufflebeam

THE AUTHORS

Daniel L. Stufflebeam retired from Western Michigan University (WMU) in 2007 as Distinguished University Professor, McKee Professor of Education, and founder of the Evaluation Center. He established the Evaluation Center at The Ohio State University (OSU) in 1963, moved it to WMU in 1973, and directed it until 2002. At OSU, he developed more than one hundred standardized achievement tests, including eight forms of the GED tests, and created the context, input, process, and product (CIPP) evaluation model. At WMU, he founded the Joint Committee on Standards for Educational Evaluation, chaired it through 1988, and led the development of standards for program and personnel evaluations. He also established and directed the national Center for Research on Educational Accountability and Teacher Evaluation and, more recently, designed WMU's interdisciplinary doctoral program in evaluation. He has received more than $25 million in grants and contracts; conducted evaluations in education, community development, housing, military personnel evaluation, state and national assessment, and other areas; lectured and consulted in twenty countries; and advised and assisted many organizations in the United States and abroad. For eight years he served on the U.S. Government Accountability Office Advisory Council on Government Auditing Standards. His publications include twenty-four books and monographs and about a hundred journal articles and book chapters. He received the American Evaluation Association (AEA) Paul Lazarsfeld prize for contributions to evaluation theory and standards, WMU's Distinguished Faculty Scholar Award, the inaugural CREATE Jason Millman Award, and membership in The Ohio State University College of Education and Human Ecology Hall of Fame. A recent book, edited with Thomas Kellaghan of Ireland, is the *International Handbook of Educational Evaluation* (2003). Stufflebeam holds a bachelor's degree from the University of Iowa and master's and PhD degrees from Purdue University.

Chris L. S. Coryn is the director of the Interdisciplinary PhD in Evaluation (IDPE) program and an associate professor in the Evaluation, Measurement, and Research (EMR) program at WMU. He received a BA in psychology in 2002 and an MA in social psychology in 2004, both from Indiana University (IU). He earned his PhD in evaluation in 2007 at WMU under Dr. Michael Scriven. He has published more than ninety peer-reviewed papers, book chapters, and monographs and is currently the executive editor of the *Journal of MultiDisciplinary Evaluation*. He has led numerous research studies and evaluations across several substantive domains, including research and evaluation in the arts and humanities, education, science and technology, health and medicine, community and international development, and social and human services. He has given lectures, speeches, and workshops nationally and internationally.

His awards include American Educational Research Association Distinguished Scholar of Research on Evaluation Award, WMU Emerging Scholar Award, AEA Marcia Guttentag Award, Michigan Association for Evaluation John A. Seeley Friend of Evaluation Award, IU Award for Graduate Research Excellence, IU Student Mentor Academic Research Team Merit Award for Outstanding Research, and IU James R. Haines Award for Outstanding Research in Psychology.

INTRODUCTION

We have planned and developed this book to aid and enlighten those who evaluate, or intend to evaluate, programs, as well as those administrators and other evaluation stakeholders who use evaluation to meet program improvement and accountability needs. The book is intended particularly for use by practicing evaluators and students in graduate programs focused on evaluation theory and practice, but its handbook nature should prove useful to evaluation clients and others with an interest in learning about evaluation and obtaining sound, effective evaluation services.

Evaluation studies should be directed toward helping clients and other stakeholders use findings well and particularly toward improving and certifying the value of evaluation services. This is a heavy professional responsibility. In this book we have drawn together information from the evaluation literature and a wide range of practical experiences to guide, to advise, and to demonstrate that success in the worthwhile pursuit of systematic evaluation is both essential and clearly possible.

Evaluation is a vital component of the continuing health of organizations. If evaluations are conducted well, organizations and their people will have the satisfaction of knowing with confidence which elements are strong and where changes are needed. Evaluation is therefore a constructive pursuit.

This book is designed as a textbook for graduate courses concerned with the critical analysis and application of program evaluation theory, approaches and models, and methods, and more widely as a handbook for use in planning, conducting, and assessing program evaluations. The book builds and expands on the widely circulated *Evaluation Models* monograph in *New Directions for Evaluation* (Stufflebeam, 2001b).

Throughout this book, we typically refer to evaluation *approaches* (rather than *models*), using the more generic term to cover all generalized ways of designing and conducting evaluations. We selected this term because it encompasses illicit as well as commendable ways of doing evaluations and includes all good approaches, whether or not they are referred to as models.

We undertook this writing project at the urging of a number of colleagues and representatives of Jossey-Bass, initially seeking only to update *Evaluation Models* (Stufflebeam, 2001b). Leaders at the publishing company convinced us, however, of the need for an updated, extended treatment of *Evaluation Models* plus practical guidelines and procedures for applying the best evaluation approaches. In this book we address these needs and also discuss the foundational topic of evaluation theory. Readers will find checklists for guiding such core evaluation tasks as designing, budgeting, contracting, reporting on, and assessing evaluations, plus others focused

on data collection and analysis. Although the heart of the book is an updated, expanded treatment of evaluation approaches (found in Parts Two and Three), this core content is now embedded in a broader discussion of theoretical and practical topics. We have focused the book on helping evaluators and others strengthen their theoretical understanding and working knowledge of evaluation.

Changes to the First Edition

This second edition of *Evaluation Theory, Models, and Applications* has undergone substantial revision since the first edition was published (Stufflebeam & Shinkfield, 2007). Major changes are the inclusion of several additional evaluation approaches and the elimination of others (and the addition of a new second author due to Anthony Shinkfield's busy schedule). In the first edition, twenty-six unique evaluation approaches were introduced and described. In this edition, the descriptions of evaluation approaches have been reduced to twenty-three, but with the addition of transformative evaluation, participatory evaluation, customer feedback evaluation, and meta-analysis. In addition, many of the evaluation approaches originally described in the first edition have been substantially revised and updated, and in many instances an approach's description has been extended to provide greater depth, detail, and insight into its specific characteristics.

Also, each chapter in this second edition begins with a list of chapter learning objectives. Key references have been added throughout chapters so that interested readers may locate additional information concerning the topic under discussion. Further, each chapter now includes a short section titled "Suggested Supplemental Readings," in which readers may locate additional source documents, books, articles, and reports intended to supplement and sometimes elaborate further on the chapter's core content. The book is also supported by relevant materials housed on the Web sites of Western Michigan University's Evaluation Center (www.wmich.edu/evalctr/) and Jossey-Bass (www.josseybass.com/go/evalmodels).

Intended Audience

Because program evaluation is such a pervasive concern in society, we have designed the book to serve the needs of a broad range of individuals and groups that must use evaluations to assess, ensure, or improve the quality of programs. The book can be useful to graduate students, evaluation and research instructors, evaluators, program administrators, business leaders, specialists in research and evaluation methodology, professionals, and other service providers who must meet requirements for public accountability, as well as those who commission program evaluations. The book treats program evaluation across disciplines, and thus is intended for use in such fields as nursing, community development, housing, education, medicine, psychotherapy, disease control, business administration, jurisprudence, national defense, engineering, social services, philanthropy, and international development, among others.

Overview of the Book's Contents

Evaluators and users of evaluations can use this book to acquire knowledge of approaches that are available for evaluating programs; the concepts and theories undergirding different evaluation approaches; and principles, standards, and procedures for guiding and judging the work of evaluators. The book provides evaluations of twenty-three evaluation approaches, detailed information about six evaluation approaches, techniques for carrying out the full range of steps in any program evaluation, and guidance for institutionalizing and mainstreaming evaluation.

Faced with a growing number of program evaluation approaches, evaluators need competence to assess and choose wisely among available options and then confidently and effectively apply the selected approach. Overall, in choosing topics for this book, we sought to provide a sense of the general nature of program evaluation, an overview and comparative analysis of alternative approaches to evaluation, in-depth instruction—with examples—in each of six ways to conduct credible program evaluations, standards for choosing among approaches, and practical guidelines for designing and carrying out an evaluation from beginning to end.

Two dominant factors—the theoretical and practical essentials of evaluation—intertwine throughout the book, and are underlined by nine themes. The first theme is:

The evaluation discipline should be grounded in sound theory—that is, a coherent set of conceptual, hypothetical, pragmatic, and ethical principles forming a general framework to guide the study and practice of evaluation.

The second theme is:

Society needs and is using evaluations to inform decisions and hold service providers accountable for the implementation and outcomes of the services they provide.

The evaluator must plan, develop, and deploy a distinctive evaluation methodology that is technically sound and responsive to the client's needs.

Part One of the book introduces program evaluation in three chapters that set out evaluation's fundamentals. Chapter 1 discusses the role of evaluation in society; defines evaluation and other key evaluation concepts; denotes the principal uses of program evaluations; identifies different, complementary methodological approaches; and describes the evaluation profession in its historical context. In general, this opening chapter offers a sweeping perspective on the evaluation field and background information for use in studying the ensuing chapters. Chapter 2 looks closely at the nature of evaluation theory, particularly program evaluation theory. It defines evaluation theory, distinguishes between evaluation models and evaluation theories, identifies criteria for judging theories, and lists illustrative hypotheses for research on program evaluation. Stressing that nothing is as useful as a sound theory, the chapter calls for increased and improved efforts to generate and validate program evaluation theories. Chapter 3 reviews and discusses principles and standards for use in guiding and assessing program evaluations. It begins with a discussion of a professionally generated set of standards for educational program evaluations (Joint Committee on Standards for Educational Evaluation, 2011) that require evaluations to meet conditions of utility, feasibility, propriety, accuracy, and

evaluation accountability. The chapter subsequently summarizes and discusses the American Evaluation Association's guiding principles for evaluators (2004), which are focused on ensuring that competent evaluators will serve the general and public welfare by conducting evaluations that are methodologically sound, competently conducted, ethical, respectful of involved and affected persons, and in the public interest. The chapter concludes with a discussion of the government auditing standards of the U.S. Government Accountability Office (2007), which cover program evaluation, as well as financial auditing, across the full range of government programs and services.

In Part Two, readers are aided in identifying, analyzing, and judging twenty-three approaches thought to cover most legitimate as well as illegitimate program evaluation efforts. This part is keyed to the book's third, fourth, and fifth themes. The third theme is:

Evaluators and clients must guard against the use of unsound, often corrupt inquiry approaches that masquerade as sound evaluation but, in fact, are designed to mislead right-to-know audiences or prevent some of them from obtaining evaluation findings.

Readers will learn to discriminate between six illicit approaches, termed "pseudoevaluations," and seventeen legitimate approaches, which are divided into four categories: quasi-evaluation, improvement- and accountability-oriented evaluation, social agenda and advocacy evaluation, and eclectic evaluation. The fourth theme is:

Evaluators can choose from a range of defensible evaluation approaches.

Because no evaluation approach is always best, the analysis in Part Two is designed to assist evaluators in choosing that one approach or combination of approaches that best fits a particular evaluation assignment. Part Two concludes with a consumer report assessment of nine of the most promising or likely to be used evaluation approaches against the requirements of the Joint Committee's *Program Evaluation Standards* (1994, 2011). Our assessments in this part are keyed to the book's fifth theme:

Evaluators should employ professional standards to assess and select evaluation approaches and ensure the quality of particular evaluations.

Part Three extends application of the fifth them by presenting detailed analysis of several widely discussed and used evaluation approaches: the experimental and quasi-experimental design approach; the case study approach, Daniel Stufflebeam's context, input, process, and product (CIPP) model; Michael Scriven's consumer-oriented evaluation approach; Robert Stake's responsive evaluation approach; and Michael Patton's utilization-focused evaluation approach. Some of these are included because we judge them to be the best available approaches. Others are discussed in depth because of their importance in the history of evaluation, as well as the likelihood of their continued use. In Part Three, readers will learn the backgrounds and orientations of the evaluation leaders who authored each approach, the approach's theoretical and philosophical underpinnings, pertinent evaluation methods and tools, and illustrations of its use.

Part Four addresses the book's sixth and seventh themes. The sixth theme is:

Evaluators should employ systematic procedures that possess general applicability across evaluation approaches and provide sound protocols for proceeding through an evaluation's

start-up, design, budgeting, contracting, information collection, analysis, synthesis, reporting, and follow-up stages.

The seventh theme is:

Evaluators should involve stakeholders in the evaluation process to hear and consider their inputs and enhance prospects for their appropriate and beneficial use of findings.

Basically, Part Four offers practical assistance, guidelines, and checklists for applying any defensible approach to evaluation. We offer down-to-earth procedures that are applicable to any sound evaluation approach. We discuss and provide illustrations of how to carry out a sequence of essential evaluation tasks: identifying, addressing, and assessing evaluation opportunities; designing, budgeting, and contracting evaluations; collecting, analyzing, and synthesizing information; and reporting and facilitating appropriate use of findings. In explaining and illustrating these tasks, Part Four emphasizes that all aspects of an evaluation must satisfy the requirements of credible standards.

Part Five consists of two capstone chapters for rounding out the book's discussion of program evaluation. Chapter 25 stresses the importance of evaluating evaluations. The chapter notes that such metaevaluations should be grounded in sound standards for evaluations, conducted formatively to guide and ensure the quality of evaluations, and conducted summatively to judge the evaluation at hand in terms of such factors as utility, feasibility, propriety, accuracy, and accountability. Chapter 25 stresses the book's eighth theme:

As professionals, evaluators must subject their evaluations to metaevaluation.

The book's final chapter turns from the previous chapters' emphasis on ad hoc program evaluation to the importance of organization-wide systems of ongoing evaluation. This concluding chapter discusses an organization's need to institutionalize systematic evaluation and mainstream its use throughout the organization. The chapter offers practical advice for designing, staffing, installing, and operating sound evaluation systems. This concluding chapter addresses the book's ninth theme:

Organizations of all types should institutionalize and mainstream sound evaluation practices as a vital part of planning programs, conducting the programs, and meeting requirements for accountability, because at its core, every discipline and service area needs sound evaluation to confirm and continually strengthen its claim that it is effectively serving clients and the public interest as well as fulfilling other defensible purposes.

All chapters conclude with review questions and one or more group exercises to help readers check and increase their understanding of the material. The book is also complemented by WMU's Evaluation Center's Web site (www.wmich.edu/evalctr/), which includes checklists for guiding evaluations according to different approaches, illustrative evaluation reports, information about evaluation training opportunities, and topical papers. From the Web site readers can also go to the open-access *Journal of MultiDisciplinary Evaluation*. The Web site is an invaluable reservoir of information of relevance to this book. Readers can greatly enhance their understanding of this book's contents by regularly consulting and making good use of tools and information on the Evaluation Center's Web site.

An instructor's supplement is also available at www.josseybass.com/go/evalmodels. Comments about this book are invited and can be sent to researchmethods@wiley.com.

Study Suggestions

As already noted, this book is a textbook, but its basic design is that of a handbook. You can turn to any chapter, independent of others, to obtain information on a particular topic. The book may be studied in groups or independently. It can be worked through from beginning to end, or its chapters can be used selectively as handbook chapters. In our desire to serve graduate education, we have sought to provide sufficient content to support a sequence of two three-semester-hour courses.

Part One is oriented to providing readers with an in-depth understanding of the evaluation discipline. The material in Part Two is especially useful in making choices among alternative approaches to evaluation, because it provides comparative analyses of different approaches. When you need to gain in-depth knowledge of a selected evaluation approach—that is, one that is already being applied in evaluating a given program or one you have selected after a review of alternatives—we advise you to see if that approach is discussed in Part Three. If it is, you can benefit by studying the pertinent Part Three chapter. Approaches not treated in Part Three can be studied in depth by consulting the suggested readings provided at the ends of the chapters. When you need practical suggestions for planning and carrying out the various steps involved in applying any evaluation approach, we suggest that you consult the procedure-oriented chapters in Part Four. The final chapters in Part Five offer detailed guidance for conducting metaevaluations and institutionalizing and mainstreaming evaluation in an organization.

We have a word of advice for graduate students and evaluation researchers who are seeking topics for research projects on evaluation. We suggest that you carefully review Chapter 2 on evaluation theory and consider designing and conducting studies to test hypotheses such as those found in that chapter. Also, you could make valuable contributions by conducting and publishing case studies on applications of given evaluation approaches discussed in Part Three or comparative studies of two or more approaches. Other useful research and development projects could entail validation of selected evaluation procedures presented in Part Four (such as the traveling observer technique or the feedback workshop technique) and development and validation of new evaluation checklists (for example, for evaluations in certain fields, such as parks and recreation, consumer products evaluation, organizational development, restaurant management, amusement parks, zoos, mail delivery, and foster care services). Finally, consider conducting and publishing metaevaluations.

A companion Web page for the book can be found at www.josseybass.com/go/evalmodels. From the book's home page readers can find additional resources, information, and materials intended to supplement this book, including unpublished chapters, course syllabi structured around the book, PowerPoint summaries of each chapter, additional study questions and exercises, checklists, spreadsheets, and links to other relevant Web sites. In addition, a companion glossary, in which terms used throughout the book's chapters are defined, can be found at the end of the book.

Summary

We offer the following steps to guide your study of this book:

- Read Chapter 1 to gain an overview of the evaluation field, an introduction to program evaluation, a historical perspective on the development of the program evaluation area, and sources of information about evaluation.
- Read Chapter 2 to gain a perspective on the importance and status of theory in the program evaluation field and on the work needed to study and improve evaluation theories.
- Read Chapter 3 to develop familiarity with the principles and standards for guiding and assessing evaluations.
- Study the chapters in Part Two to distinguish proper from improper evaluations, identify the range of available creditable evaluation approaches, and see a consumer report evaluation of selected approaches.
- In Part Three, develop in-depth knowledge of any of the six evaluation approaches provided.
- Study the chapters in Part Four to identify practical procedures for carrying out the various steps in an evaluation.
- Consult the final two chapters in Part Five for guidance and practical tools for evaluating evaluation plans and reports and for ideas and tools of use in helping an organization develop and employ a system of ongoing evaluation.
- For the book as a whole and for each chapter, list what you see as your high-priority learning objectives, keep notes on your progress in achieving the objectives, and ultimately evaluate your learning gains.
- After reading each chapter, respond to the review questions and one or more group exercises to make sure you have grasped the chapter's main points, and then reread chapter material as appropriate.

We recommend that after reading the book, you try your hand at applying your new knowledge by designing evaluations of programs of interest; evaluating published evaluations; teaching others about evaluation theory, approaches, and procedures; and designing and conducting research on evaluation. Good luck!

PART ONE

FUNDAMENTALS OF EVALUATION

Part One provides information on the foundations of evaluation. In the following three chapters we give an overview of the evaluation field, analyze the state of theory in the field, and describe the field's guiding principles and standards. These chapters afford an appreciation of the history and status of the evaluation discipline and address some of the key theoretical and professional issues facing its theoreticians and practitioners.

CHAPTER 1

OVERVIEW OF THE EVALUATION FIELD

Evaluation is perhaps society's most fundamental discipline; it is an essential characteristic of the human condition; and it is the single most important and sophisticated cognitive process in the repertoire of human reasoning and logic (Osgood, Suci, & Tannenbaum, 1957). It permeates all areas of human activity and has important implications for maintaining and improving services and protecting citizens in all areas of interest to society. Evaluation is a process for giving attestations to such matters as reliability, effectiveness, cost-effectiveness, efficiency, safety, ease of use, and probity. Society and individual clients are at risk to the extent that services, products, and other objects of interest are of poor quality. Evaluation serves society by providing affirmations of worth, value, progress, accreditation, and accountability—and, when necessary, a credible, defensible, nonarbitrary basis for terminating bad programs or, conversely, expanding good programs.

What Are Appropriate Objects of Evaluations and Related Subdisciplines of Evaluation?

In general, we refer to objects of evaluations as evaluands. When the evaluand is a person, however, we follow Scriven's recommendation to label the person whose qualifications or performance is being evaluated as the evaluee (Scriven, 1991). Objects of evaluations may be programs, projects, policies, proposals, products, equipment, services, concepts and theories, data and other types of information, individuals, or organizations, among others. Although the practice of evaluation largely concentrates on

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- The distinction between formal and informal evaluation
- The potential contributions and limitations of formal evaluation
- Evaluation as a profession and its relationship to other professions
- Conceptual and operational definitions of evaluation
- Key criteria for evaluating programs, including merit and worth
- The roles of values clarification and setting standards in reaching evaluative conclusions
- Four main uses of evaluation
- Distinctions between formative evaluation and summative evaluation
- Distinctions between research and evaluation
- Historical milestones in the development of professional evaluation

program evaluation, one can refer to a range of other areas of evaluative inquiry, such as personnel evaluation, product evaluation, portfolio evaluation, performance evaluation, proposal evaluation, and policy evaluation. The scope of evaluation applications broadens greatly when one considers the wide range of disciplines, activities, and endeavors to which evaluation applies. One can speak, for example, of educational evaluation, social and human services evaluation, arts evaluation, consumer product evaluation, human resources development and evaluation, city planning and evaluation, real estate appraising, engineering testing and evaluation, hospital evaluation, drug testing, manufacturing evaluation, science policy evaluation, evaluation of international development and international aid, agricultural experimentation, and environmental evaluation.

Are Evaluations Enough to Control Quality, Guide Improvement, and Protect Consumers?

The presence of sound evaluation does not necessarily guarantee high quality in services or that those in authority will heed the lessons of evaluation and take needed corrective actions. Evaluations provide only one of the ingredients needed for quality assurance and improvement. There are many examples of defective products that have harmed consumers not because of a lack of pertinent evaluative information, but because of a failure on the part of decision makers to heed and act on rather than ignore or cover up alarming evaluative information. The continued sales of the Corvair automobile after its developers and marketers knew of its rear-end collision fire hazard provides one clear example (see also Nader, 1965). Here we see that society has a critical need not only for competent evaluators but for evaluation-oriented decision makers as well. For evaluations to make a positive difference, policymakers, regulatory bodies, service providers, and others must obtain and act responsibly on evaluation findings. The production and appropriate use of sound evaluation constitute one of the most vital contributors to strong services and societal progress.

Evaluation as a Profession and Its Relationship to Other Professions

As a profession with important roles in society, evaluation has technical aspects requiring thorough and ongoing training. It possesses an extensive and rapidly developing professional literature containing information on evaluation models and methods and findings from research on evaluation (Christie, 2011; Coryn & Westine, 2013). Its research material evolves from, and is closely connected to, the wide range of evaluations conducted in all fields. Evaluation has many professional organizations, including the American Evaluation Association (AEA) and other state and national evaluation associations. Among the earliest known professional societies were the May 12th Group, Division H of the American Educational Research Association (AERA), the Evaluation Network (E-Net), and the Evaluation Research Society (ERS), all of which originated in the late 1960s and early 1970s.

In 1995 there were only five evaluation organizations worldwide, including AEA (ensuing from the merger of E-Net and ERS in 1986), the Canadian Evaluation Society (CES), the Australasian Evaluation Society (AES), the European Evaluation Society, and the Central American Evaluation Society. By 2006 there were more than fifty national and regional evaluation organizations throughout the world, most in developing countries (Segone & Ocampo, 2006). There are also university training programs in evaluation, among them the Interdisciplinary PhD in Evaluation (IDPE) program and the Evaluation, Measurement, and Research (EMR) program at Western Michigan University (Coryn, Stufflebeam, Davidson, & Scriven, 2010), as well as other evaluation graduate programs at Claremont Graduate University, the University of Illinois, The Ohio State University, the University of Minnesota, the University of North Carolina, the University of Virginia, and the University of California at Los Angeles (for historical trends in graduate training in evaluation, see LaVelle and Donaldson [2010]). In addition, the field has developed recognized standards for evaluation services, including the Joint Committee on Standards for Educational Evaluation's standards for evaluating programs, personnel, and students (1981, 1988, 1994, 2003, 2009, 2011) and the U.S. Government Accountability Office's *Government Auditing Standards* (U.S. General Accounting Office, 2002; U.S. Government Accountability Office, 2003, 2007), plus AEA's *Guiding Principles for Evaluators* (2004).

To communicate and disseminate developments in, thinking about, and critiques of evaluation theory, methods, and practice, professional journals and other types of publications dedicated exclusively to evaluation scholarship and practice began to appear in the 1970s (Coryn, 2007a). One of the field's earliest publications, which first appeared in 1974, was the journal *Evaluation and Program Planning*. This was followed in 1975 by the journal *Studies in Evaluation;* in 1976 by *Evaluation Review: A Journal of Applied Social Research;* some years later by the *American Journal of Evaluation* (formerly published under the titles *Evaluation News,* prior to 1986, and *Evaluation Practice,* between 1986 and 1997); *New Directions for Evaluation* (formerly *New Directions for Program Evaluation*) and *Evaluation & the Health Professions,* both of which appeared in 1978; and *Educational Evaluation and Policy Analysis,* which first appeared in 1979.

The 1980s were marked by the appearance of the *Canadian Journal of Program Evaluation,* which emerged in 1986; the *Journal of Personnel Evaluation in Education* (now published under the title *Educational Assessment, Evaluation, and Accountability*), which was first published in 1987; and *Practical Assessment, Research and Evaluation,* which was launched in 1988.

In the 1990s several additional journals appeared, including *Research Evaluation* in 1991, which is published in the Netherlands; *Evaluation: The International Journal of Theory, Research and Practice,* which is published in the United Kingdom; and the *Journal of Evaluation in Clinical Practice,* the last two having first been published in 1995. In the next decade several more scholarly journals devoted to evaluation emerged, including the *Evaluation Journal of Australasia,* which was first published in 2000, and the *Journal of MultiDisciplinary Evaluation,* which first appeared in 2004.

Despite the burgeoning number of scholarly evaluation journals, many evaluation scholars and practitioners disseminate their work in discipline-specific journals, including those found

in education, health and medicine, philosophy, psychology, and sociology, to name but a few. In addition to publishing in evaluation and discipline-specific journals, other evaluation scholars publish their work in subject-specific areas, such as measurement, research, and statistics.

As a distinct profession, evaluation is supportive of all other professions and in turn is supported by many of them; no profession could excel without evaluation. Services and research can lead to progress and stand up to public and professional scrutiny only if they are regularly subjected to rigorous evaluation and shown to be sound. Also, improvement-oriented self-evaluation is a hallmark of professionalism. Program leaders and all members of any profession are obligated to serve their clients well. This requires that they regularly evaluate, improve, and be accountable for their contributions. In the sense of assessing and improving quality and meeting accountability requirements, all professions (including evaluation) are dependent on evaluation. Moreover, evaluation draws concepts, criteria, and methods from such other fields as philosophy, political science, psychology, sociology, anthropology, education, economics, communication, public administration, information technology, statistics, and measurement. Clearly it is important for evaluators to recognize and build on the symbiotic relationships between evaluation and other fields of study and practice.

Improvements in programs and other evaluands can be enhanced and made more enduring to the extent that supporting evaluations are relevant, systematic, rigorous, and timely, and to the extent that clients make responsible use of findings. Evaluations that lack these aspects of discipline typically are fruitless, wasteful, and misleading. It bears mention, however, that evaluators can only do their best, and despite strenuous efforts to involve clients in evaluations, there is no certainty that clients will heed and act on sound evaluation findings. If rigorous evaluations are to make a positive difference, clients must play their part by helping focus evaluations, supporting their conduct, and making sound use of findings. Accordingly, evaluation training programs should prepare evaluation specialists *and* evaluation clients to collaborate effectively in conducting evaluations that are both rigorous and useful.

What Is Evaluation?

Mainly because there have been different approaches to evaluation over the years, definitions of the term *evaluation* have themselves varied. In earlier times, for example, evaluation was commonly associated with assessing achievement against clearly defined objectives, or (in schools and universities) conducting norm-referenced testing, or (in such fields as agriculture and experimental psychology) conducting controlled experiments. Also, particularly during the 1970s, many evaluations were keyed only or mainly to professional judgment. Subsequently, there was a growing belief that useful evaluations are ones that provide quality information for making and assessing decisions. These and other concepts of evaluation have elements of credibility, depending often on the type of evaluation study being undertaken and especially the needs of the evaluation users.

One of the earliest and still most prominent definitions of evaluation states that it means determining whether objectives have been achieved. Although following this definition can guide one to assess accomplishments in achieving one's valued goals, in a broader

sense the practice of objectives-based evaluation has serious limitations and can even be counterproductive. Especially from the perspective of an independent evaluation of consumer products or services, employing the objectives-based evaluation approach can cause an evaluation to fail. One of this approach's problems is that some objectives are unworthy of achievement. Surely evaluators must avoid judging a program as successful solely because it achieved its own objectives. Objectives might well be corrupt, dysfunctional, unimportant, not oriented to the needs of intended beneficiaries, or mainly reflective of a developer's profit motive or other conflicts of interest. Another problem is that this approach steers evaluations in the direction of looking only at outcomes. Many evaluations also should examine a program's objectives, structure, and processes, especially if the evaluation is to contribute to program improvement or adoption and adaptation by other service providers. Moreover, a focus on objectives might cause evaluators not to search for important, unintended consequences (often called side effects). These can be beneficial or harmful, as is often seen in prescription drugs that may do as much harm as good for particular users. In addition to the deficiencies already noted, evaluators employing an objectives-based evaluation approach provide feedback only at the completion of a program. Depending on the needs of the client group, evaluators often should also deliver timely findings for use in planning and in guiding programs toward successful outcomes.

Definitions that equate evaluation with any one methodology should be rejected. Sometimes evaluations based on randomized experiments can provide consumers with useful information on the comparative outcomes of competing programs, products, or services. However, in many evaluations, a controlled experimental approach would not be feasible, or it would be counterproductive; it might be unethical; or it might fail to address key questions about needs, objectives, plans, processes, side effects, and other important aspects of a program. Similarly, other useful methods—such as sample surveys, standardized testing, site visits, or self-studies—are far too narrow in the information they yield to provide a sufficient basis for most program evaluations. Evaluation, therefore, rather than being equated with any one methodology, should encompass all methods that are necessary and useful to reach defensible judgments of programs or other entities, and evaluators should selectively apply appropriate methods.

In this book we advocate a basic definition of evaluation put forth by the Joint Committee in 1994.[1] We present three variations of the definition. First, we present the definition as the Joint Committee stated it.[2] The committee's definition is general, calling for evaluations to be systematic and focused on determining an object's value. We then extend the general definition to highlight a range of important, generic criteria for consideration when assessing programs. Finally, we expand the definition further to outline the key steps involved in carrying out a sound evaluation and to stress the importance of obtaining both descriptive and judgmental information. We see the Joint Committee definition as especially appropriate and useful when conversing with uninitiated audiences and focusing their attention on the essence of evaluation. The second rendition can be helpful when discussing with clients or other stakeholder groups the values that should be referenced when evaluating a particular program or other object. The third version is especially appropriate when planning the required evaluation work.

Joint Committee Definition of Evaluation

The Joint Committee's 1994 definition states that "evaluation is the systematic assessment of the worth or merit of an object" (p. 3). Advantages of this definition are that it is concise and consistent with common dictionary meanings of evaluation. We see this as the definition to use when discussing evaluation at a general level. Notably, some alternative definitions of evaluation often also include significance, resulting in a formal definition of evaluation as the act or process of determining the merit, worth, or significance of something or the product of that process (Davidson, 2005; Scriven, 1991).

Evaluation's root term, *value,* denotes that evaluations essentially involve making value judgments. Accordingly, evaluations are not value-free (Scriven, 1993). They need to reference pertinent values. Depending on the particular program or other evaluand, such values may include effectiveness, efficiency, usability, cost, safety, legality, and so on. Also, the evaluation itself should be grounded in some defensible set of values for judging evaluations. Here we see that an evaluation is an evaluand that should adhere to relevant values for judging evaluations. These may include professionally defined principles (as in AEA's *Guiding Principles for Evaluators* or the Joint Committee program evaluation standards). Essentially, an evaluation—be it an assessment of a program or of an evaluation—should assess the evaluand's standing against the referenced values. This truism presents evaluators with the impetus to choose appropriate values for judging an evaluand. For example, in evaluating public services in the United States, evaluators should be true to, and sometimes specifically invoke, such democratic precepts as freedom, equity, due process of law, and the need for an enlightened society. Moreover, as will be explained in Chapter 3, evaluators should hold their evaluations to meeting such values as the Joint Committee–defined standards of utility, feasibility, propriety, accuracy, and evaluation accountability.

The Joint Committee's 1994 definition partially addresses the need to determine values by denoting that evaluations should assess merit or worth. Scriven (1991) pointed to the nontrivial differences between these two concepts and their important role in determining an evaluand's value. According to both Scriven (1991) and the Joint Committee (1994), merit essentially involves excellence or quality (that is, intrinsic value), whereas worth includes merit within the context of a particular culture and its associated needs, costs, and related circumstances (that is, extrinsic value). In Table 1.1, the essential characteristics and nature of these concepts are summarized, with further discussion following.

Merit

In general, one needs to look at the merit or quality of an evaluand. For example, does a state's special program for preparing middle school history teachers succeed in producing teachers who confidently and effectively teach middle school students about pertinent areas and periods of history? In general, does an evaluand do well what it is supposed to do? If so, it rates high on merit. The criteria of merit reside in the standards of the evaluand's particular discipline or area of service. In the example here, an evaluator might base her or

Table 1.1 Characteristics of Merit and Worth

Merit	Worth
May be assessed on any object of interest	Is assessed only on objects that have demonstrated an acceptable level of quality
Pertains to the intrinsic value of the object	Pertains to the extrinsic value of the object
Pertains to quality, that is, an object's level of excellence	Pertains to an object's quality and value or importance within a given context
Is assessed using the question, Does the object do well what it is intended to do?	Is assessed using the question, Is the object of high quality and also something a target group needs?
Is tied to accepted standards of quality for the type of object being evaluated	Is tied to accepted standards of quality and to data from a pertinent needs assessment
Concerns the object's rating on standards of quality and against competitive objects of the same type	Entails judgments of the object's quality and importance and value to a particular consumer group
May be assessed through comparison of an object with standards or competitive objects	Assessments of worth may be comparative or noncomparative

his assessment of merit on published standards of effective teaching and the state's required content for middle school history programs. Graduates of the program would thus be assessed on knowledge of the required history content and effectiveness in teaching the content. The subject program would be judged high on merit to the extent that graduates scored high on pertinent measures of content knowledge and teaching competence. Merit (or quality) then can be broadly understood as intrinsic excellence in the absence of costs.

Worth

An evaluand that rates high on merit might not be worthy. By worth, we refer to an evaluand's combination of excellence and service in an area of clear need within a specified context and considering the costs involved (both monetary and nonmonetary). Suppose the middle school program is a special emergency program developed and funded at a previous time when the state's colleges and universities were graduating too few history teachers to meet the needs of schools in the state. Suppose further that more recently, the state's universities have increased their production of competent middle school history teachers, and many of these new teachers cannot find jobs. Arguably, the state no longer needs the special emergency program, because the state's universities are now supplying more qualified middle school history teachers than the schools can employ. In this situation, although the state's special program has good merit, it now has low worth to the state and does not warrant continued investment of the state's scarce resources. We see in this example that this high-quality program's worth could be gauged only after an assessment of the need for the program's graduates. Here, we see that assessments of worth have to be keyed to assessments of need within the context of a particular setting and time period. Broadly, then, worth (or value) is quality under consideration of context and costs.

Needs

By a need, we refer to something that is necessary or useful for fulfilling a defensible purpose, without which satisfactory functioning cannot occur. We define a defensible purpose as a legitimately defined, desired end that is consistent with a guiding philosophy, set of professional standards, institutional mission, mandated curriculum, national constitution, or public referendum, for example. Other terms to describe defensible purposes are *legitimatized mandates, goals,* and *priorities*. In the middle school illustration, presumably the state curriculum requires that all students in the state be well educated in designated areas of history. This "defensible purpose" requires further that school districts employ competent history teachers. In this case, a competent history teacher fits our definition of an entity that is necessary or useful for fulfilling the defensible purpose of sound history instruction—that is, a need. Because of the state's finding that this need is now being fulfilled by state colleges and universities, this excellent special program would now meet the criterion of merit but not the criterion of worth. In reaching judgments of something's worth, evaluators should identify needs, then determine whether they are being met, partially met, or unmet in the context of interest (Stufflebeam, McCormick, Brinkerhoff, & Nelson, 1985; also see Coryn, Gugiu, Davidson, & Schröter, 2008).

Needs may be of either the outcome or the treatment variety (also see Davidson, 2005). An outcome need is a level of achievement or outcome in a particular area required to fulfill a defensible purpose, such as preparing students for higher education. For example, high school students need to develop competencies in mathematics, science, social studies, and language arts to enter top-notch colleges and universities. A treatment need is a certain service, service provider, or other helping agent required to meet an outcome need. To continue the example, a school district needs an appropriate curriculum and competent teachers (the treatment needs) to help students attain areas and levels of competence (the outcome needs) required for admission to high-level colleges and universities. One assesses both treatment and outcome needs to determine whether they are being met or unmet and whether they are consonant.

Typically, the meeting of outcome needs depends on meeting the treatment needs. For example, a dentist would be likely to check patients with tooth decay for use of fluoridated water or toothpaste. Here the outcome need (for cavity-free teeth) is not being met, and it is prudent to check on the treatment need related to fluoridation. In contrast, if patients evidence no tooth decay, the dentist would be unlikely to check them for use of fluoridated water or toothpaste.

Needs Assessments

In general, a needs assessment is a systematic investigation of the extent to which treatment and/or outcome needs are being met (Stufflebeam, McCormick, et al., 1985). One might posit that comprehensive high schools should serve the defensible purpose of developing students in all areas of human growth and development: intellectual, psychological, social, physical, moral, vocational, and aesthetic. In an appropriate range of curricular areas, a comparison of students' scores on standardized achievement tests to criterion-referenced standards or norms would give an indication of whether students' intellectual outcome needs were being met. However, considering the school's intention to develop students also in physical, aesthetic, psychological,

Table 1.2 Concepts Related to Needs and Needs Assessment

Concept	Definition	Example
Defensible purpose	A desired end that has been legitimated	Students' development of basic academic skills
Need	Something that is necessary or useful for fulfilling a defensible purpose	Competent, effective instruction in the basic skill areas
Outcome need	An achievement or outcome required to meet a defensible purpose	Students' demonstration of proficiency in specified areas, such as twelfth-grade math, science, and language arts
Treatment need	A certain service, competent service provider, or other helping agent	Competent instructors in twelfth-grade courses in math, science, and language arts
Needs assessment	A systematic investigation of the extent to which treatment and/or outcome needs are being met	Examination of students' scores on national tests and evaluation of the involved teachers

social, moral, and vocational areas, the achievement test scores would be insufficient to assess the full range of questions concerning students' outcome needs. To be valid, needs assessments have to be keyed to the full range of intended outcomes.

Some needs assessments will have a narrow scope and appropriately address a quite restricted construction of outcome needs. Even in a narrowly focused program, however, it can be important to consider a broad range of outcome and associated treatment needs. For example, school-based instrumental music programs contribute to students' development in such areas as social relations, psychological well-being, discipline, and employment. In general, an assessment of a program's worth should assess and gauge its quality and outcomes against the assessed outcome and treatment needs of beneficiaries. Table 1.2 offers a summary of key concepts related to needs and needs assessment.

Evaluations Should Be Systematic

Beyond its focus on merit and worth, the Joint Committee's 1994 definition of evaluation requires evaluations to be systematic. We acknowledge that the broad meaning of evaluation encompasses haphazard or unsystematic evaluations as well as carefully conducted evaluations. In this book, we are advocating for and discussing the latter. Indeed, this book is intended as a countermeasure to careless or corrupt inquiry processes that masquerade as evaluations and often lead to biased or otherwise erroneous interpretations of something's value. Instead, we seek the kind of evaluation that is conducted with great care—not only in collecting information of high quality but also in clarifying and providing a defensible rationale for the value perspectives used to interpret the findings and reach judgments and in communicating evaluation findings to the client and other audiences.

An Extended, Values-Oriented Definition of Evaluation

Although the Joint Committee's 1994 definition of evaluation has the positive features just noted, it omits mention of other key generic values. We thus extend the definition of evaluation as follows: evaluation is the systematic assessment of an object's merit, worth, probity, feasibility,

safety, significance, and/or equity. We see the values referenced in this definition as particularly important in a free and democratic society, but also acknowledge that we might have included additional values. Of course, evaluators have to engage in a good deal of values clarification as they plan their studies. Those included in our extended definition of evaluation are a good set to consider, but evaluators and their clients often should invoke additional values that pertain to the contexts of particular studies and the unique cultures and interests of stakeholders. Nonetheless, many sound and defensible evaluations will be strongly influenced by some or all of the five values we have added to merit and worth. In the following paragraphs we discuss and elucidate each of the values noted in the extended definition of evaluation.

Probity

During the writing of the first edition of this book (Stufflebeam & Shinkfield, 2007), there was a rash of public scandals in which major corporations based in the United States defrauded shareholders and others out of billions of dollars. Moreover, at least one major audit firm that contracted to evaluate a corporation's financial conditions and lawful operations was found to have complicity in that corporation's fraud. This audit firm compromised its independence and credibility. Not only did it fail to report on the probity of the corporation's accounting practices, but also it was alleged to have distorted and covered up information to hide the company's unethical, unlawful practices. Here we see that the corporation cheated its shareholders, workers, and ultimately the public, and that the audit firm was charged with aiding and abetting the fraud. On another front, there have been despicable scandals across the globe in which clergy and teachers have been found to be pedophiles.

Clearly, the public interest (broadly defined) requires that evaluations address considerations of probity: assessments of honesty, integrity, and ethical behavior. Unless there is no prospect for fraud or other illicit behavior, evaluators should check on a program's uncompromising adherence to moral standards. However, when probity breaches are expected, there is cause to err on the side of too much consideration of probity in evaluations of programs and institutions. To the extent required to form a defense against unethical behavior, probity considerations should be addressed in many evaluations of programs and in evaluations of evaluations.

Feasibility

Although a program (or service, or other type of service-oriented evaluand) might be of high quality, directed to an area of high need, and unimpeachable on ethical grounds, it still could fail on the criterion of feasibility. For example, it might consume more resources than required or cause no end of political turmoil. If either is the case, the program should at least be modified in these areas to make it more feasible. Obviously a good evaluation of the program should speak to this issue and, where appropriate, provide direction for making the program easy to apply, efficient in the use of time and resources, and politically and culturally viable. Evaluation of a program's feasibility sometimes justifies a cancellation decision. This argument in favor of assessing feasibility seems applicable to all programs (and to all service-oriented evaluands).

Safety

Many evaluations focus squarely on the issue of safety. Obvious cases are evaluations of new pharmaceutical products, medical treatments, laboratory equipment, meat and other food products, automobiles, railroad transportation services, air traffic control, oil and gas production and distribution, stepladders, electrical equipment, children's toys, and insecticides. Consumers are at risk to the extent that such commodities and services are manufactured, sold, and dispensed or delivered without rigorous safety checks and appropriate cautions. Moreover, many programs also require evaluations that examine the safety of facilities, equipment, activity regimens, crowd control practices, and others. To see the importance of safety evaluations in programs, one need only recall head injuries in football, lost teeth in ice hockey, heat strokes in a variety of outdoor sports, fires and explosions in school laboratories, fires resulting in many deaths due to improper fire escapes or fire drills, and fatalities due to faulty school buses or incompetent bus drivers. The criterion of safety applies to evaluations in all fields and to evaluations of programs as well as of products and services.

Significance

Another criterion that sometimes comes into play is a program's significance: its potential influence, importance, and visibility. Many programs are of only local or short-term interest. Other programs that have far-reaching implications should be examined and judged on the significance of their mission and outcomes. Such an assessment can be especially important in deciding whether and how far to disseminate lessons learned and in helping interested parties make sound decisions concerning adopting, adapting, and/or disseminating all or particular aspects of a program. Evaluators should consider the possibility that the program under study has far-reaching implications outside the local arena and possibly should be evaluated for its significance over time and in other settings.

Equity

The last generic evaluative criterion to be mentioned here is equity, which is predominantly tied to democratic societies. It argues for equal opportunities for all people and emphasizes freedom for all (also see House & Howe, 2000a). In the United States, an educational evaluation of a public educational service would be incomplete if it did not assess whether the service is provided for, and made available to, public school students from all sectors of society. This concept of equity is complex. It is not enough to say that public educational services may be sought and used by all people. As Kellaghan (1982) has argued, for example, when there is true equity in education, there will be seven indications of its existence:

1. A society's public educational services will be provided for all people.
2. People from all segments of the society will have equal access to the services.
3. There will be close to equal participation by all groups in the use of the services.
4. Levels of attainment—for example, years in the education system—will be substantially the same for different groups.

5. Levels of proficiency in achieving all of the education system's objectives will be equivalent for different groups.
6. Levels of aspiration for life pursuits will be similar across societal groups.
7. The education system will make similar impacts on improving the life accomplishments of all segments of the population (especially ethnic, gender, and socioeconomic groups) that the educational system serves.

We assert that equity, in the broadest sense, is an important criterion for all evaluations that involve delivering programs to groups of people.

Operationalizing Our Definition of Evaluation

The extended definition of evaluation has provided an expanded look at key generic criteria for evaluating programs. From the discussion, it is evident that the Joint Committee's 1994 definition of evaluation and our adaptation focused on generic evaluative criteria are deceptive in their apparent simplicity. When one takes seriously the root term *value,* then inevitably one must consider value perspectives of individuals, groups, and organizations, as well as information. The combining of these in efforts to reach determinations of the value of something cannot be ignored. To serve the needs of clients and other interested persons, the information supplied to support evaluative judgments should reflect the full range of appropriate values.

We now expand the definition to outline the main tasks in any program evaluation and denote the types of information to be collected. Our operational definition of evaluation states that evaluation is the systematic process of delineating, obtaining, reporting, and applying descriptive and judgmental information about some object's merit, worth, probity, feasibility, safety, significance, and/or equity. One added element in this definition concerns the generic steps in conducting an evaluation. The other new element is that evaluations should produce both descriptive and judgmental information.

It is important to note that the work of evaluation includes both interface/communication and technical tasks. In regard to the interface aspects, evaluators communicate with clients and other stakeholders in the interest of planning relevant evaluations; conveying clear, timely findings; and assisting with use of the findings. To ensure an evaluation's relevance and impact, the evaluator needs to effectively engage stakeholders in the evaluation's planning and use. The technical tasks are concerned with the research aspects of an evaluation: the collection, organization, analysis, and synthesis of information. Evaluators need to be competent in both the communication and technical aspects of evaluation (also see Stevahn, King, Ghere, & Minnema, 2005). This competence is best acquired through formal courses and experiences in planning, conducting, and reporting on a wide range of evaluations. We have characterized the work of evaluation in four tasks: delineating, obtaining, reporting, and applying. Part Four of this book addresses these process tasks in detail.

Delineating

The delineating task entails the evaluator's interacting with the client and other program stakeholders. The aim here is to focus the evaluation on key questions, identify key audiences,

clarify pertinent values and criteria, determine information requirements, project needed analyses, construct an evaluation budget, and effect contractual agreements to both govern and facilitate the evaluation work. Basically, the delineating task encompasses effective, interactive communication involving evaluator, client, and other interested parties and culminates in negotiated terms for the evaluation. Particular areas of needed expertise include audience analysis, listening, developing rapport, interviewing, situational and cultural analysis, values clarification, conceptualization, proposal development, negotiation, contracting, and budgeting. The results of these actions should set the stage for the ensuing data collection work. In fact, delineating activities extend throughout the evaluation in response to the program's changing circumstances, identification of new audiences, continuing interaction with stakeholders, and emerging information needs. Moreover, a delineation process that is carried out thoroughly and professionally establishes a basis for essential trust and rapport between an evaluator and a client group.

Obtaining

The obtaining task encompasses all of the work involved in collecting, correcting, organizing, analyzing, and synthesizing information. Key areas of required expertise are research design, sampling, measurement, interviewing, observation, site visits, archival studies, case studies, focus groups, photography, database development and management, statistics, content analysis, cost analysis, policy analysis, synthesis, and computer technology. Program evaluators need expertise in these and related technical areas to provide clients with sound, meaningful, and creditable information. Results of the obtaining work are grist for preparing and presenting oral and printed evaluation reports.

Reporting

In the reporting task, the evaluator provides the client and other audiences with feedback. Typically such work includes preparing and delivering interim oral and printed reports, multimedia presentations, press releases, printed final reports, and executive summaries, as well as ongoing informal exchanges with the evaluation's client and, often, stakeholders. The point of all such reporting activities is to communicate effectively and accurately the evaluation's findings in a timely manner to interested and right-to-know audiences and to foster effective uses of evaluation findings. Reporting activities, in various forms, occur throughout and after completion of an evaluation (Coryn, 2006). Particular areas of needed expertise are writing, formatting reports, editing, information technology, oral communication, leading of group discussions, and dissemination. Effective reporting sets the stage for applying the evaluation findings.

Applying

The applying task is under the control of the client and other users of the evaluation. Nevertheless, the evaluator should at least offer to assist in the application of findings. Such assistance might be follow-up workshops, a critique of the client group's plans to apply findings, coordination of focus group deliberations, or responses to questions from the client

or other users. We have found that clients appreciate this kind of assistance from evaluators. It is seen as a continuation of the evaluation itself, provided that the initiative comes from the client after the evaluator offers this "rounding-off" service. Assisting in the sound use of evaluation findings requires forethought and funding. In starting an evaluation, therefore, the evaluator and client should consider the possibility of the evaluator's involvement in the application stage and should plan, budget, and contract for such follow-up assistance as appropriate. To be effective in supporting the application of evaluation findings, evaluators need to be knowledgeable about principles and procedures of effective change and research on evaluation use (see also Alkin, Daillak, & White, 1979; Patton, 1997, 2008). Also, they need skills in the areas of communication, consulting, group process, and counseling (see also Dewey, Montrosse, Schröter, Sullins, & Mattox, 2008).

Descriptive and Judgmental Information

The final major feature of our operational definition of evaluation concerns the nature of information included in evaluations. From experience, we know that sound, useful evaluations are grounded in descriptive and judgmental information. In general, audiences for evaluation reports want to know what program was evaluated, how well it was carried out, and how good it was, requiring the evaluator to collect and report both descriptive and judgmental information.

Descriptive Information A final evaluation report should describe a program's goals, plans, funding, staffing, operations, and outcomes objectively (that is, as factual statements). As much as possible, the descriptive information should be kept separate from judgments of the program. Relatively pure, dispassionate descriptions of a program are needed to help evaluation audiences know, for example, what the evaluated program was like, how it was staffed and financed, how it operated, how much time was required for implementation, how much it cost, and what would be required to replicate it. The evaluator also has a vested interest in getting a clear view of the program apart from how other observers judged it. This is especially important when interpreting a program's outcomes and judging its success. For example, in judging the effects of a community's immunization program on childhood diseases, an evaluator needs to determine and report the extent to which the pertinent inoculations were administered to all the targeted children as planned. If they were not, the deficient outcome more likely is due to poor program implementation than defects in the program plan.

Judgmental Information Beyond the collection of descriptive information, it is equally important to gather, assess, and synthesize judgments of a program. According to the values-oriented definition of evaluation given earlier, sound evaluations involve judging an evaluand against a set of values. Values-oriented feedback can be a vital, positive force when it is integral to development, directed toward identifying strengths as well as weaknesses, focused on improving the evaluand, and grounded in evidence or at least experience with the program. Appropriate sources of judgments include program beneficiaries, program staff, pertinent experts, and (of course) the evaluator, among others. Such judgments are typically reached

through the integration or synthesis of facts (that is, descriptive information) and values, or the synthesis of multiple statements of value (Coryn, 2007; Davidson, 2005; Scriven, 1991, 1993).

How Good Is Good Enough? How Bad Is Intolerable? How Are These Questions Addressed?

Many evaluations carry a need to draw a definitive conclusion or make a definite decision on quality, safety, or some other variable. For example, funding organizations regularly have to decide which proposed projects to fund, basing their decisions on these projects' relative quality, costs, and importance compared with other possible uses of available funds (also see Coryn, Hattie, Scriven, & Hartmann, 2007; Coryn & Scriven, 2008; Scriven & Coryn, 2008). For a project already funded, the funding organization often needs to determine after a funding cycle whether the project is sufficiently good and important to continue or increase its funds. In trials, a court has to decide whether the accused is guilty or not guilty. In determinations of how to adjudicate drunk-driving charges, state or other government agencies set decision rules concerning the level of alcohol in a driver's blood that is legally acceptable. These examples are not just abstractions. They reflect true, frequent circumstances in society in which evaluations have to be definitive and decisive.

The problem of how to reach a just, defensible, clear-cut decision never has an easy solution. In a sense, most protocols for such precise evaluative determinations are arbitrary, but they are not necessarily capricious. Although many decision rules are set carefully in light of relevant research and experience or legislative processes, the rules are human constructions, and their precise requirements arguably could vary, especially over time. The arbitrariness of a cut score (for example, a score that classifies scores above it [the cut line] as good and those below it as unsatisfactory) is also apparent in different α (alpha) and β (beta) levels that investigators may invoke for determining statistical significance. Typically, α is set, by convention, at 0.05 or 0.01, but it might as easily be set at 0.06 or 0.02. In spite of the difficulties in setting and defending criterion levels, societal groups have devised workable procedures that more or less are reasonable and defensible for drawing definitive evaluative conclusions and making associated decisions. These procedures include applying courts' rules of evidence and engaging juries of peers to reach consensus on a defendant's guilt or innocence; setting levels for determining statistical significance and statistical power; using fingerprints and DNA testing to determine identity; rating institutions or consumer products; ranking job applicants or project proposals for funding; applying cut scores to students' achievement test results; polling constituents; grading school homework assignments; contrasting students' tested performance with national norms; appropriating and allocating available funds across competing services; and charging an authority figure with deciding, or engaging an expert panel to determine, a project's future. Although none of these procedures is beyond challenge, as a group they have addressed society's need for workable, defensible, nonarbitrary decision-making tools (also see Cizek & Bunch, 2007).

Some of these procedures have in common the advance setting of cut scores, standards, or decision rules. In the United States, for example, it is known in advance that all twelve

(or sometimes six) members of a jury must vote "guilty" for a defendant in a criminal trial to be found guilty beyond a reasonable doubt. Advance determinations of criteria and acceptable levels also apply to evaluations of new drugs; drunk-driving convictions; and certification of safe levels in water, air quality, food products, and bicycle helmets, for example.

When it is feasible and appropriate to set standards, criterion levels, or decision rules in advance, a general process can be followed to reach precise evaluative conclusions. The steps suggested here would be approximately as follows: (1) define the evaluand and its boundaries; (2) determine the key evaluation questions; (3) identify and define crucial criteria of goodness or acceptability; (4) determine as much as possible the rules for answering the key evaluation questions, such as cut scores and decision rubrics; (5) describe the evaluand's context, cultural circumstances, structure, operations, and outcomes; (6) take appropriate measurements related to the evaluative criteria; (7) thoughtfully examine and analyze the obtained measures and descriptive information; (8) follow a systematic, transparent, documented process to reach the needed evaluative conclusions; (9) subject the total evaluation to an independent assessment; and (10) confirm or modify the evaluative conclusions.

Although this process is intended to provide rationality, rigor, fairness, balance, and transparency in reaching evaluative conclusions, it rarely is applicable to most of the program evaluations treated in this book. This is so because often one cannot precisely define beforehand the appropriate standards and evaluative criteria, plus defensible levels of soundness for each one and for all as a group. So how do evaluators function when they have to make plans, identify criteria, and interpret outcomes without the benefit of advance decisions on these matters? There is no single answer to this question. More often than not, criteria and decision rules have to be determined along the way. We suggest that it is often best to address the issues in defining criteria through an ongoing, interactive approach to evaluation design, analysis, and interpretation and, especially, by including the systematic engagement of a representative range of stakeholders in the deliberative process.

What Are Performance Standards? How Should They Be Applied?

Often evaluation is characterized as comparing a performance to a standard (see also Fournier, 1995; Scriven, 1980). Constructing or setting performance standards is the process of setting one or more cut scores against which the performance of something is judged, with the cuts score(s) representing two or more states, conditions, or degrees of performance. Cut scores divide a distribution of performances into two or more discrete categories. So, for example, in the case where only a single cut score is set, its application results in the creation of only two possible performance categories, such as pass or fail (for example, for issuing a driver's license or a license to practice medicine or law). In some contexts, however, multiple cut scores may be required, the application of which results in the creation of more than two performance categories. Here, such cut scores are exemplified by the method used in a typical grading system (that is, A, B, C, D, or F) or that used for the National Assessment of Educational Progress (that is, advanced, proficient, or basic). Such classification methods can most easily be described as approaches that are either norm referenced or criterion referenced.

Norm-referenced methods include those whereby a distribution of a norm group's scores on a variable of interest is established and used to determine where the score of a separate evaluand places in the normative distribution, such as how far the obtained score is above or below the mean of the normative distribution's scores. Such norms-based conclusions typically are expressed in the evaluand's percentile rank against the normative distribution's table of scores. In contrast to norm-referenced methods of standard setting, and more commonly used, are criterion-referenced methods. With criterion-referenced methods of standard setting, performance does not depend on how well other objects perform. Criterion-referenced methods are also sometimes referred to as absolute methods, in contrast to the relativistic nature of norm-referenced methods.

The concept of criterion-referenced assessment is perhaps clearest in the judging of livestock, cats, and dogs, where associations of breeders publish the standards for particular breeds. Similarly, the sports of diving, gymnastics, and figure skating have published standards against which to judge performances by athletes. However, observers often view with disdain the lack of transparency, reliability, and validity of rendered judgments. The problems are even more acute in most standards-based program evaluations in which there are no juried, published standards for particular classes of programs. In such cases, evaluators and clients often have to concoct and agree on standards by which to judge particular programs.

Sometimes clients and their evaluators define behavioral objectives that, among other things, specify cut scores for distinguishing good performance from poor performance on each variable of interest. Many problems follow from this practice. The objectives are arbitrary and often unrealistic. They may not reflect the assessed needs of the intended beneficiaries. They may be more appropriate for average performers than for very high or low performers. For example, beneficiaries who already far exceed the cut score standard may find a disincentive for improvement in the program's low expectations. At the other end of the distribution, beneficiaries who are far below the standard may believe it is futile to attempt to reach the cut score standard, and may consequently give up. Also, cut score standards have a tendency to narrow a program's focus; lock it into predetermined objectives; and inhibit it from responding over time to emergent needs, developments, insights, and opportunities to exceed past performance.

An alternative to this narrow, preordinate approach to standards-based evaluation (*preordinate* being a coined term common in evaluation circles signifying rigid, advance stipulation of an evaluation's questions, standards for interpreting findings, and measurement and analysis procedures) is to view the evaluation process as a flexible, creative, evolving, responsive approach to assessing and supporting the client group's continuing quest for program improvement. W. Edwards Deming (see Walton, 1986) sold a similar notion to Japanese automobile manufacturers in the 1970s and helped spawn an amazing trend of continuing improvement in the quality of automobiles that eventually spread throughout the world. Deming's notion was not to attain and continue to achieve at any given level of quality, but continually to strive for better and better quality. Moreover, the recommended focus was on continuously improving the quality of manufacturing processes under the assumption (which proved true) that this

would result in both improved manufacturing processes and better and better outcomes. In the education field, W. L. Sanders and Horn (1994) argued similarly that the standard for educational programs should be continued growth and improvement for every student, whatever her or his prior level of achievement.

We believe it makes no sense to close the gap between high and low achievers, because sound education that helps all students reach their fullest potential will inevitably widen the achievement gap. This claim can be rejected only if one also rejects the claim that humans vary in abilities and capacities. To do the latter would require discarding society's huge store of evidence from research on individual differences.

Why Is It Appropriate to Consider Multiple Values?

Many evaluations face the challenge of multiple value perspectives. This is part and parcel of the world's increasingly pluralistic societies. Addressing competing and often conflicting values and cultures of different members of an evaluation audience is a necessary and difficult task in evaluations (also see Shadish, Cook, & Leviton, 1991). We would argue that it is the shared and differential needs of the consumers of a given service that should be ascertained as a basis for determining what information to collect and what standards to invoke in determining the worth of that service.

Sometimes an evaluator should address the value conflict issue by separately interpreting process and outcome information against the distinct sets of values or priorities held by different segments of the stakeholder population. Moreover, the evaluator might beneficially seek out and assess alternative programs or services to determine which ones best meet the needs of different stakeholder groups.

In planning evaluations, evaluators should deal directly with the important matter of choosing and applying pertinent values (also see Scriven, 1994b, 2007). They should determine what sets of values will be referenced in interpreting findings and sometimes in searching for and analyzing program options. Such determinations require evaluators to work within their basic philosophical convictions—that is, to act with integrity. Evaluators also should take into account a program's mission and the pertinent cultures, values, needs, and priorities of the program's leaders as well as impactees and other stakeholder groups. In issuing evaluative conclusions or putting forward assessments of alternative programs, evaluators should report the employed values and explain why they were chosen.

Addressing conflicting values is not an easy task for evaluators, if for no other reason than that they are not the sole arbiters of one set of values over another. Our advice is, first, never to take the side of one group rather than another's and, second, to take a dispassionate view of the needs of differing value groups and work toward the formulation of a sound set of guiding values that reflects integrity and the interests of the different parties to the evaluation. That being said, evaluators should not set aside their basic values, such as those concerning human rights. They should not proceed with an evaluation if doing so would aid and abet unethical or immoral decisions and actions. Clearly, an evaluator should decline an evaluation assignment if it is alien to his or her beliefs about what is sound, moral behavior.

Should Evaluations Be Comparative, Noncomparative, or Both?

Evaluators may focus on a single product or service or compare it with alternatives. Depending on the circumstances, an evaluation legitimately may be comparative or noncomparative. A main consideration is the nature of the audience and what evaluative information it needs. If the audience is composed of consumers who need to choose a product or service, the evaluation should be comparative and help consumers learn what alternatives are available and how they compare on critical criteria. If the audience includes developers or consumers who are already committed to the development or use of a given program, the evaluation might focus intensively on the workings of the program and help provide direction for improving it. Periodically, however, even if a group is firmly devoted to a certain service or product, it might get a better version from the provider of this service or product or find a better alternative by opening consideration to other providers, or by engaging in a systematic process of invention and innovation.

In general, we think that evaluations should be comparative before the purchase of a product or service or the beginning of a program, noncomparative during program development or use of a service, and periodically comparative after development or sustained use to open the way for improvements or better alternatives. Whether an evaluation should be comparative depends on the intended uses of the evaluation. If, for example, a selection is to be made from among alternative programs or uses of resources, then the evaluation should clearly be comparative.

How Should Evaluations Be Used?

We see four main uses of evaluations: improvement, accountability, dissemination, and enlightenment.

Formative Evaluations for Improvement

The first use is to provide information for developing a service, ensuring its quality, or improving it. Evaluations to serve this use typically are labeled formative evaluations (Scriven, 1967). Basically, they provide feedback for improvement. They are prospective and proactive. They are typically conducted during development of a program or its ongoing operation. Formative evaluations offer guidance to those who are responsible for ensuring and improving the program's quality and who should, in doing so, pay close attention to the nature and needs of the program's consumers. In formative evaluations, evaluators assess and assist with the formulation of goals and priorities, provide direction for planning by assessing alternative courses of action and draft plans, and guide program management by assessing implementation of plans and interim results.

Information from a formative evaluation is directed toward improving operations, especially those that are in the process of development. In the main, formative evaluations serve quality assurance purposes. In formative evaluations, the evaluator should interact closely with program staff and provide guidance for decision making. The evaluation plan needs to be

flexible and responsive. When the main aim is to improve an existing program, the evaluation should resemble a case study more than a comparative experiment. In fact, locked-in, controlled experiments that require random assignment of program participants to alternative program treatments and keeping treatments stable and unchanging typically prevent the evaluator from giving to program personnel the ongoing feedback for improvement that is the essence of formative evaluations.

Summative Evaluations for Accountability

The second main use of evaluations is to produce summative reports (Scriven, 1967). These are retrospective assessments of such evaluands as completed projects, established programs, finished products, or services rendered. Summative evaluations typically occur following development of a product, completion of a program, or end of a service cycle. They draw together and supplement previously collected information and provide an overall judgment of the evaluand's value. Summative evaluations are useful in ascertaining accountability for successes and failures, informing consumers about the quality and safety of products and services, and helping interested parties increase their understanding of the assessed phenomena. Summative evaluation reports are not aimed primarily at the development staff, but at the sponsor and consumers. The reports should convey a cumulative record of what was done and accomplished and an assessment of the evaluand's cost-effectiveness. Information derived from in-depth case studies and field tests is of interest to audience members in such situations. Field tests can involve productive use of comparative experiments. In the medical field, for example, results from double-blind studies comparing a newly developed treatment or other evaluand to a placebo or another competitive treatment can help potential users decide whether to use the new contribution. Whereas in general we argue against the use of experimental design in formative evaluations, it can be useful in some summative evaluations. This is especially the case in evaluations designed to undergird dissemination of a final product, service, program, project, or other evaluand. But even then, a randomized experiment is only part of a sound summative evaluation.

Relationship Between Formative and Summative Evaluations

Table 1.3 summarizes main features of formative evaluation and summative evaluation.

Both formative and summative evaluations are needed in developing and certifying evaluands, including programs, projects, products, or services; or, in the case of personnel, they help in developing potential and gauging the extent to which required criteria for certification, tenure, promotion, and the like are met. Too often, only summative evaluation is carried out—for judging an evaluand's past performance. This restricts development processes and may lead to inadequate or even incorrect conclusions. Subjecting a trainee nurse to an accountability assessment, for example, while ignoring the obvious advantages of fostering improvement through formative methodologies is foolish. Similarly, when a new model of an agricultural combine is being designed and developed, a lack of formative information covering cost, efficiency, reliability, safety, ease of use, durability, effectiveness, and potential marketing

Table 1.3 Formative Evaluation and Summative Evaluation

Descriptor	Formative Evaluation	Summative Evaluation
Purpose	Quality assurance; improvement	Providing an overall judgment of the evaluand
Use	Guiding decision making	Ascertaining accountability for successes and failures; promoting understanding of assessed phenomena
Functions	Provides feedback for improvement	Informs consumers about an evaluand's value (for example, its quality, cost, utility, competitive advantage, and safety)
Orientation	Prospective and proactive	Retrospective and retroactive
When conducted	During development or ongoing operations	After completion of development
Particular types of services	Assists with goal setting, planning, and management	Assists consumers in making wise decisions
Foci	Goals, alternative courses of action, plans, implementation of plans, interim results	Completed projects, established programs, or finished products; ultimate outcomes; costs; side effects
Variables	All aspects of an evolving, developing program	A comprehensive range of dimensions having to do with merit, worth, probity, safety, equity, and significance
Audience	Managers, staff; connected closely to insiders	Sponsors, consumers, and other interested stakeholders; projected especially to outsiders
Nature of evaluation plans	Flexible, emergent, responsive, interactive	Relatively fixed, not emergent or evolving
Typical methods	Case studies, observation, interviews (controlled experiments typically are inappropriate here)	A wide range of methods, including case studies, controlled experiments, and checklists
Nature of reports	Periodic, often relatively informal, responsive to client and staff requests	Containing a cumulative record and assessment of what was done and accomplished, a comparison between the evaluand and critical competitors, and a cost-effectiveness analysis
Relationship between formative evaluation and summative evaluation	Often forms the basis for and supplements summative evaluations	Involves compiling, assessing, and building on previously collected formative evaluative information

would be disastrous for the manufacturers. Evaluations delayed until the near completion of an employee's training and probationary period, or a product's development, or a project or program's implementation, or a service's full period of delivery may be too late to foster needed improvements and produce successful outcomes.

The relative emphases of formative and summative evaluations will change according to the nature of and circumstances surrounding the evaluand. In general, as portrayed in Figure 1.1, formative evaluation will be dominant in a program's early stages and less so as the program matures. Summative evaluation will take over as the program concludes and certainly will be used after it is completed. All concerned in these evaluations should have a clear understanding of when and in what circumstances formative evaluation may give way to summative evaluation. The conclusion should not be drawn, however, that all evaluations fall

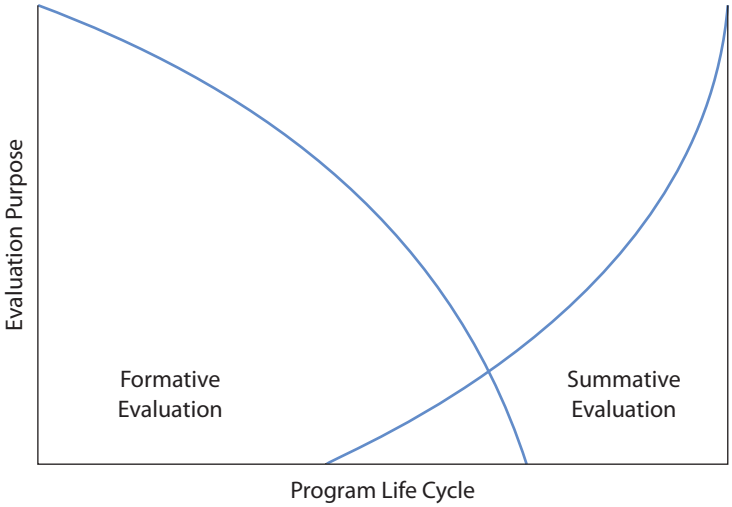

Figure 1.1 Relationship Between Program Life Cycle and Evaluation Purpose

into one or both categories. Many of the evaluation approaches depicted later in this book can be used for formative purposes, summative purposes, or both. Moreover, additional purposes (such as program monitoring) have been put forth, and have been widely debated (for example, Chen, 1996; Patton, 1996; Scriven, 1993, 1996; Wholey, 1996), yet we believe that formative and summative evaluations adequately reflect the large majority of work that evaluators engage in.

Stake (1969) made an interesting observation, apropos the relationship between formative and summative evaluations, that formative evaluations are closely connected to "insiders"—that is, program developers—whereas summative evaluations are of more interest to "outsiders"—that is, the potential users of the developing (or developed) programs. This does not assume that formative evaluations are necessarily undertaken by internal personnel or that summative evaluations are always conducted externally. A wide array of factors, such as timelines, finances, and the competency of personnel to undertake evaluations, will often determine whether evaluations, either formative or summative, are internal or external. The dominant question to be answered is whether the process and findings are credible. Ideally, internal summative evaluations are subjected to external audits (see Chapter 25 on metaevaluations).

Finally, formative evaluations often form the basis for summative evaluations. If this is to occur, both the evaluators and those who commission the studies must agree and also make clear to all involved that a formative evaluation will be conducted and used to form the basis for a subsequent summative evaluation. It should also be recognized that on occasion, the soundness and utility of a formative evaluation may be strengthened by the intervention of interim summative evaluations (usually carried out by external personnel) at critical points of a program's development. Such interplay between separately conducted formative and summative evaluations requires sound professional collaboration, which is a hallmark of good evaluation practice. (For other dimensions of this topic, see "Why Are Internal Evaluation Mechanisms Needed?" later in this chapter.)

Evaluations to Assist Dissemination Efforts

The third use of evaluations is to help developers disseminate proven practices or products and help consumers make wise adoption or purchasing decisions. Here the evaluator must critically compare the service or product with competitors. Perhaps the best example of evaluations aimed at serving dissemination and informing adoption decisions are those found in *Consumer Reports*. Each issue of this well-known monthly magazine provides independent evaluations of alternatives for consumer products and services: automobiles, insurance policies, mortgages, breakfast cereals, chain saws, refrigerators, computers, cameras, cell phones, restaurant chains, supermarket chains, hotel chains, and house paints, to name just a few. The unique feature of evaluations for dissemination is their focus on questions of practical interest to consumers. In Parts Two and Three, we describe Michael Scriven's consumer-oriented evaluation approach, which is predominately premised on a product model of evaluation.

A more recent example is the numerous evidence-based repositories, including the Campbell Collaboration, Cochrane Collaboration, and What Works Clearinghouse, among many others, that have become increasingly commonplace in the last few decades as mechanisms for using evaluation results to disseminate effective practices or products (Coryn, Tarsilla, & Hobson, 2010). These repositories largely provide information about the results of meta-analyses and randomized controlled trials in health and medicine, human and social services, and education. In part, these repositories emerged due to the climate of increasingly scarce resources and greater demands for accountability, in which policymakers and those in practice-based disciplines and professions have been seeking high-quality, nonarbitrary, and defensible evidence for formulating, endorsing, and, occasionally, enforcing best policies and practices (Flay et al., 2005).

Evaluations to Foster Enlightenment

The fourth use of evaluations is to foster enlightenment, or new understandings arising from evaluations (also see Chelimsky, 1997; Patton 1997, 2008). Basically, evaluation and research are different enterprises. Evaluators attempt to consider all criteria that apply in determining value, whereas researchers may be restricted to the study of selected variables that are of interest in testing theory, diagnosing problems, or answering particular questions. Evaluations typically involve subjective approaches and are not as tightly controlled and subject to manipulation as is the typical research investigation. However, efforts over a period of time to evaluate a program or set of similar programs may produce information of use in evolving and testing theory. Certainly evaluation results often should and do lead to focused, applied research efforts and sometimes to development of institutional or social policies. Hence, we believe that in planning studies, evaluators should consider how their findings might contribute to new insights in matters of interest to theorists, policymakers, and scientists, especially through formulation of testable hypotheses. With some forethought, careful planning, and appropriate budgeting, evaluations may serve not only to guide operating programs, sum up and assess their contributions, and lead to the dissemination of effective products and services but also to address particular research, theory, or policy questions.

Why Is It Important to Distinguish Between Informal Evaluation and Formal Evaluation?

By this point, it should be clear that program evaluation is a demanding field of practice. At the same time, everybody essentially evaluates constantly, whether making choices about the trivial or the critical (see also Posavac & Carey, 2003). We believe it is important to distinguish formal evaluation from informal evaluation. In fact, the distinction is at the root of the need for and emergence of the evaluation profession. Just as most individuals employ home remedies and over-the-counter medications in addressing their minor ailments, almost everybody recognizes that some health issues require diagnosis and treatment by competent physicians in accordance with the standards of the medical profession. Similarly, many evaluations can and must be conducted on an informal basis, whereas others require a rigorous, systematic approach, often including an independent perspective.

Informal Evaluations

Everybody performs informal evaluations whenever judging and making decisions about the things observed, thought about, interacted with, or being considered for purchase. For example, we do this when purchasing food, cars, tools, refrigerators, computers, computer programs, stocks, correspondence courses, insurance policies, or termite protection services. Depending on the nature of the evaluand, one might look for options, read labels, consult friends who have pertinent experience, form a committee or task group to deliberate on the evaluative questions of interest, call the Better Business Bureau, consult other consumer information sources, search the Internet for consumers' assessments of items purchased, or try out something before deciding to keep it. These are all good and appropriate evaluative moves and fit within our general conception of informal evaluation. The conduct of informal evaluations, however, is prone to haphazard data collection, crediting and using propaganda and other forms of misinformation, errors of judgment, strong influence by salespersons, acting on old preferences or prejudices, relying on out-of-date information, depending on an inadequate or biased sample of customer feedback, or making expedient choices. In many cases, the steps in an informal evaluation are unsystematic, lacking in rigor, and based on biased perspectives. Thus, informal evaluations typically offer a weak basis for convincing decision makers and others of the validity of evaluation findings and the appropriateness of ensuing conclusions and recommendations. We can get by with weak, informal evaluations when only we have to pay the price and abide by the consequences. Better, more formal evaluations are called for when there is a need to inform critically important decisions, especially ones that will affect many people, require substantial expenditures, or pose substantial risk.

Formal Evaluations

In accordance with the definition of evaluation given earlier, formal evaluations should be systematic and rigorous. By the term *systematic,* we refer to evaluations that are relevant, designed and executed to control bias, kept consistent with appropriate professional standards,

documented, reported to right-to-know audiences, and otherwise made useful and defensible. Especially, we define formal evaluations as ones that are held up to scrutiny against appropriate standards of the evaluation profession. The kind of formal evaluation we are promoting requires systematic effort by one or more persons who have the requisite evaluation competencies. We do not disparage the informal evaluations that are part and parcel of everybody's daily life, any more than we would advise people not to make prudent use of home remedies and over-the-counter medications. Moreover, not all formal evaluations need to be conducted by outside evaluation experts. What is required is that those conducting the evaluation meet the standards of the evaluation field. In Chapter 3 we summarize professionally developed guiding principles for evaluators (AEA, 2004); professional standards for program evaluations (Joint Committee, 1981, 1994, 2011); and the U.S. government auditing standards (U.S. General Accounting Office, 2002; U.S. Government Accountability Office, 2003, 2007). Building on these, this book is designed to help evaluation and research students, practicing evaluators, evaluation clients, research methodologists, and other interested parties attain the perspectives and basic level of proficiency required to undertake defensible formal evaluations grounded in professional standards and principles for practice.

How Do Service Organizations Meet Requirements for Public Accountability?

We cannot stress too strongly that society is dependent on sound evaluations to obtain safe, high-quality goods and services from a wide range of professionals and the organizations in which they work. Such organizations include school districts, universities, research centers, hospitals, government departments, charitable foundations, churches, community service organizations, and others. Any operatives should deliver services that are of high quality, up to date, safe, efficient, fairly priced, honest, and generally in the public interest. To meet accountability requirements, each profession, public service area, and society should regularly subject services to formal evaluations. Some evaluation work is appropriately directed at regulation and protection of the public interest. This work should be conducted by independent bodies, including government agencies, accrediting boards, and external evaluators. Equally important are the formative and summative evaluations of services that professionals and other service providers and their organizations themselves conduct. These internal or self-evaluations are an important aid to continually scrutinizing and improving services and also supplying data needed by independent or external evaluators.

Accreditation

A wide range of accrediting organizations periodically assess the performance of member organizations against formally established standards. Typical accreditation evaluations are grounded in clear accreditation criteria and guidelines for self-assessment. In the accreditation process, the institution or program to be evaluated proceeds by conducting a lengthy process of self-assessment, typically lasting at least a year. A team of external evaluators, appointed

by the accrediting organization, then reviews the self-report, conducts a site visit, and writes an independent evaluation report. The accrediting organization subsequently uses the report to make decisions on whether, to what extent, and for what period the subject institution or program is to be accredited and submits its report to the institution or program. Typically accreditation is awarded for a finite period, such as five years. The accrediting body then updates its publicly available list of accredited institutions or programs. In some cases, provisional accreditation is provided pending corrective actions by the assessed institution or program. A prime accreditation criterion often is that the subject institution or program will operate an internal evaluation mechanism and make use of its findings.

Why Are Internal Evaluation Mechanisms Needed?

Some large school districts, medical schools, foundations, and government agencies maintain well-funded and adequately staffed evaluation offices, and their evaluators have succeeded in helping their institutions be accountable to constituents, obtain guidance for planning and administering their services, win grants and contracts, and meet requirements of accrediting organizations or other oversight bodies. To keep their services up to date and ensure that they are effectively and safely meeting their clients' needs, service institutions and programs should continually obtain pertinent evaluative feedback. This process includes studying the outcome and treatment needs of their clients; evaluating relevant approaches that are being proposed or used elsewhere; evaluating the performance of personnel; closely monitoring and assessing the delivery of services; assessing immediate and long-term outcomes; and searching for ways to make services more efficient, effective, and safe. Conducting such internal evaluations is a challenging task. The credibility of internal evaluation is enhanced when it is subjected periodically to metaevaluation (Scriven, 1969b; Stufflebeam, 1978, 2001c; also see Chapter 25), in which an independent evaluator evaluates and reports publicly on the quality of internal evaluation work. Such independent metaevaluation also provides direction for strengthening the internal evaluation services. Optimally, metaevaluations are both formative and summative.

Chapter 26 provides in-depth information on how organizations may institutionalize and mainstream a systematic process of internal evaluation.

Why Is Evaluation a Personal as Well as an Institutional Responsibility?

Even if an organization has a strong internal evaluation unit, every professional in the organization needs to engage in systematic evaluation. There is no escaping the fact that evaluation is a personal as well as an organizational responsibility. Offices of evaluation and accrediting firms can help organizations meet their major responsibilities in regard to continuous evaluation and accountability. An office of evaluation can also provide an organization's staff with in-service training and technical support in evaluation. However, all professionals bear responsibility for formally evaluating their own performance. It is in their interest to do so, because evaluation is an essential means of finding out and acting on what is going right and wrong. Moreover, conducting and acting on sound evaluation constitute a fundamental part of what it means to be a professional—a member of an established profession

who continually works to deliver better services. We hope this book will both inspire and assist individual professionals and other service providers as well as evaluation students and specialists, enabling them to develop evaluation competencies and effectively carry out systematic evaluations.

What Are the Methods of Formal Evaluation?

One aspect that distinguishes formal evaluation from informal evaluation is the area of methodology. When we move our consideration away from evaluations that involve quick, intuitive judgments toward those that entail rigorously gathered findings and effective communication, we must necessarily deal with the complex areas of epistemology, rules of evidence, information sciences, research design, measurement, statistics, communication, and some others. Many principles, tools, and strategies within these areas are pertinent to systematic evaluation. The well-prepared evaluator will have a good command of concepts and techniques in all these areas and will stay informed about potentially useful technological developments. Evaluators who would exert leadership and help advance their profession should contribute to the critiquing of existing methods and the development of new ones.

Over the years, many evaluators have exclusively chosen and used—even championed—a narrow set of techniques. Some evaluators have equated evaluation with their favorite methods—for example, experimental design, standardized testing, surveying, case studies, site visits by teams of experts, or participant observation. Other leaders have sharply attacked narrow views of which methods are appropriate and argued for a broader, more eclectic approach, which is where we find ourselves (also see Mark, Henry, & Julnes, 2000). A key point in the latter position is that the use of multiple methods and perspectives enhances the dependability of inferences and conclusions and yields appropriate levels of circumspection.

We believe that evaluators should know about a wide range of pertinent techniques and how well they apply in different evaluation contexts. Then, in each evaluation, they can assess which techniques are potentially applicable and which most likely would work best and in combination to serve the given study's particular purposes. Among the technical areas in which we think the professional evaluator should be proficient are proposal writing, research design, budgeting, contracting, scheduling, system analysis, logic models, interviewing, focus groups, survey research, case studies, content analysis, observation, checklists, goal-free evaluation, advocate teams, test construction, rating scales, database development and management, statistical analysis, cost analysis, technical writing, and project administration.

What Is the Evaluation Profession, and How Strong Is It?

The formal profession of evaluation emerged only during the last third of the twentieth century. In so short a time period, this young profession has made remarkable progress, but it still has far to go. The evaluation field now has national and state professional societies of evaluators; annual conventions; a substantial literature and knowledge base, including numerous professional journals (also see Coryn, 2007) and a wide range of theoretical and

technical books; specialized Web sites; discussion groups, blogs, and listservs (also see Christie & Azzam, 2004); master's and doctoral programs (also see LaVelle & Donaldson, 2010); institutes and workshops on specialized evaluation topics (for example, the Evaluators' Institute, which presents annual training sessions in specific evaluation procedures); client organizations that fund a wide range of evaluations; evaluation companies; guiding principles for evaluators; and standards for program, personnel, and student evaluations. These are substantial gains considering the field's status in 1964, when it had none of these elements. The evaluation field is still immature, however, when compared with established professions, such as medicine, law, engineering, and accounting, and other service areas, such as those of master plumbers, licensed electricians, and dental hygienists. In particular, the evaluation field lacks some of the hallmarks of a mature profession. For example, membership in AEA is open to anyone regardless of training and expertise in evaluation. Furthermore, the field has no mechanisms for certifying or licensing competent evaluators (also see S. C. Jones & Worthen, 1999; Worthen, 1999), although the Canadian Evaluation Society began a credentialing effort for evaluators in 2010. Despite the field's substantial progress, clients of evaluation have no formal means of determining which self-proclaimed evaluators have been certified as competent. And even though evaluations are widely recognized as essential to the health of any organization, acceptance of tertiary training to gain qualifications as an evaluator is lagging worldwide. The evaluation field's stature and credibility are threatened by its lack of professional certification and quality control. This is especially so because there is "much gold in the evaluation hills," and because, in our experience, all too often ill-prepared evaluators obtain high-cost contracts to conduct evaluation assignments for which they lack the needed expertise.

What Are the Main Historical Milestones in the Evaluation Field's Development?

The evaluation field has evidenced only modest efforts to systematically record and analyze its history (for example, Shadish & Luellen, 2005). Any profession, to effectively serve its clients, must evolve in response to changing societal needs and in consideration of theoretical and technical advancements. Unless the members of a profession develop and maintain a historical perspective on their work, they are likely to persevere in using a stagnant conception of their role, not to remember valuable lessons of the past, not to stimulate and contribute to innovation in their field, and all too frequently to return to deficient methods of the past. It has been said often that those who do not learn from their history are doomed to repeat it.

In this section we focus on the history of the program evaluation field, especially as evaluation theory and practice have evolved in the area of education (Stufflebeam, Madaus, & Kellaghan, 2000).[3] We believe this is appropriate and will be instructive, because the profession of evaluation developed earliest and most heavily within the field of education. We provide only a brief historical sketch, noting the most significant developments in educational program evaluation.

Our historical analysis is grounded in the seminal work of Ralph W. Tyler (described later in this book), who is often spoken of as the father of educational evaluation. Using his

initial contributions as the main reference point, we have identified six major periods: (1) the Pre-Tylerian Period, which includes developments before 1930; (2) the Tylerian Age, which spans 1930 to 1945; (3) the Age of Innocence, which runs from 1946 to 1957; (4) the Age of Realism, which covers years 1958 through 1972; (5) the Age of Professionalism, which includes developments from 1973 to 2004; and (6) the Age of Global and Multidisciplinary Expansion, from 2005 to the present.

The Pre-Tylerian Period: Developments Before 1930

Systematic evaluation was not unknown before 1930, but it was not a recognizable movement. In the mid-1840s in the United States, the common method of assessing student learning and the quality of instruction was an annual oral examination conducted by school committees. Because of a desire for more dependable inspections of schools, in 1845 Boston replaced the oral exams with the first systematic school survey using printed tests. Horace Mann championed this approach and advised Boston to base school policies on factual results from testing the eldest class in each of the city's nineteen schools. The committee running the survey faced problems similar to those seen in today's large testing programs. In particular, teachers felt threatened because they knew that their students' test scores would be viewed as an indicator of their teaching competence.

The initial tests reflected the curriculum of the day, mainly requiring abstract renderings consistent with the prevalent Puritan philosophy. They were chalk-and-slate or quill-and-paper tests, requiring students mainly to recall facts but, in a minor way, also to demonstrate application of what they had learned. Members of the school committees administered the tests during six hours over two days. Test results overall were discouraging. Reports contained a brief, often negative evaluative statement about each school. Mann saw these new methods of inspecting schools as impartial, thorough, and accurate in assessing what pupils had been taught and lauded their use in arriving at independent judgments of schools. In today's language, we could say he judged the new evaluation approach as meeting conditions of objectivity, validity, and reliability. Although the Boston survey spawned similar examination projects elsewhere in the United States, it was not until the end of the nineteenth century that end-of-semester printed tests became a common feature in schools nationwide.

It is generally recognized that Joseph Rice conducted the first formal educational program evaluation in the United States. An education reformer who provided educational administrators in New York City with leadership, Rice in 1895 launched the most ambitious plan ever undertaken to collect data on education. His goal was to confirm that student learning was deficient. Over the next decade, he obtained test scores in spelling and mathematics from about sixteen thousand students. A key finding was that the amount of time spent on spelling each day related little to spelling achievement. The Boston and Rice surveys gave publicity to the survey technique as a means of collecting and analyzing data to help identify and correct deficiencies in the schools and form sound educational policies. Its use in the twentieth century was evident in the 1915 publication of the Cleveland Education Survey. Sponsored by the Survey Committee of the Cleveland Foundation, the twenty-five-volume report assessed every

aspect of the school system and was heralded as the most comprehensive study of an entire school system ever completed.

The dawning of the twentieth century saw the emergence of yet another approach to evaluation. In applying the concepts of efficiency and standardization to manufacturing, Frederick Taylor had found standardization to contribute to efficiency and assurance of consistent quality in manufactured products. Taylor's success in manufacturing influenced leaders in education to seek standardization and efficiency in schools. Consequently, under the leadership of Edward Thorndike and others, educators launched the now massive enterprise of standardized testing. They believed that standardized tests could check the effectiveness of education and thereby show the way to more efficient student learning. Technology for measuring student achievement and other human characteristics developed strongly in the United States, Great Britain, and some other countries throughout the twentieth century, and, in this century, continues to be developed and widely applied. Educators and the public have often looked to scores from standardized tests as a basis for judging schools, programs, teachers, and students. Nevertheless, perhaps no other educational practice has generated so much criticism and controversy as has standardized testing, especially when high stakes have been attached to the results (American Evaluation Association Task Force on High Stakes Testing, 2002).

As a countermovement to rigid testing practices, a progressive education movement developed during the 1920s that espoused the ideas of John Dewey and even earlier writers. Travers (1983) stated the matter extremely well:

> Those engaged in the progressive education movement viewed the new emphasis on standardized achievement testing as a menace to everything they hoped to accomplish. They wanted to make radical changes in the curriculum, but the standardized tests tended to encourage the retention of the established curriculum content. They wanted to emphasize the development of thinking skills, but the tests placed emphasis on the memorization of facts. They wanted to emphasize self-evaluation, with the child's own evaluation of himself as the point from which progress should be measured, but the achievement testers encouraged a competitive system in which a child was judged in terms of his position in a group. The use of criterion-referenced tests was minimal in the 1920s and 1930s, and although such tests would have answered this last criticism of the progressive educators, it would not have resolved even a small fraction of the misgivings that the progressives had about the new achievement testing. (p. 144)

Despite a continuing flow of criticism, the use of objective achievement tests has continued to expand. The limitations of tests in measuring important educational outcomes, such as abilities to understand, apply, and critique, often are discounted in favor of obtaining quick and easy measures. In the service of educational evaluation, large-scale testing programs have been extremely expensive. We also judge them as grossly inadequate for assessing programs and institutions on merit, worth, probity, feasibility, significance, safety, and equity.

Objective testing can play a useful role in educational program evaluations, but it can provide only a small part of the needed information.

Although program evaluation has only recently been identified as a field of professional practice, this account illustrates that systematic program evaluation is not a completely recent phenomenon. Some of the modern evaluation work (testing commissions, surveys, accreditation, and experimental comparison of competitors) continues to draw from ideas and techniques that were applied long ago.

The Tylerian Age: 1930 to 1945

In the early 1930s Tyler coined the term *educational evaluation* and published a broad and innovative view of both curriculum and evaluation. Over about fifteen years, he developed his ideas until they constituted an approach that provided a clear-cut alternative to other views (Madaus, 2004; Madaus & Stufflebeam, 1988).

What mainly distinguished his approach was its concentration on clearly stated objectives. In fact, he defined evaluation as determining whether objectives have been achieved. In light of this definition, evaluators were supposed to help curriculum developers clarify the student behaviors that were to be produced through the implementation of a curriculum. The resulting behavioral objectives were then to provide the basis for both curriculum and test development. Curriculum design was thus influenced away from the content to be taught and toward the student behaviors to be developed. The technology of test development was to be expanded to provide for tests and other assessment exercises referenced to objectives as well as those referenced to individual differences and national or state norms.

During the 1930s the United States, as well as the rest of the world, was in the depths of the Great Depression. Schools and other public institutions had stagnated from a lack of resources and optimism. Just as Franklin Roosevelt tried to lead the American economy out of this abyss through his New Deal program, Dewey and others tried to help education become a dynamic, innovative, and self-renewing system. Called progressive education, this movement reflected the philosophy of pragmatism and employed the tools of behavioral psychology.

Tyler was drawn into this movement when he was commissioned to direct the research component of the now famous Eight-Year Study (E. R. Smith & Tyler, 1942), which was designed to examine the effectiveness of certain innovative curricula and teaching strategies being employed in thirty schools throughout the United States. The study is noteworthy because it helped Tyler at once expand, test, and demonstrate his conception of educational evaluation.

Through this nationally visible study, Tyler was able to publicize what he saw as clear-cut advantages of his approach over others. Because Tylerian evaluation involves internal comparisons of outcomes with objectives, it does not require costly and disruptive comparisons between experimental and control groups. The approach concentrates on direct measures of achievement, as opposed to indirect approaches that measure such inputs as quality of teaching, number of books in the library, extent of materials, and community involvement. Tylerian evaluations need not be heavily concerned with reliability of differences between the scores

of individual students, and they typically cover a wider range of outcome variables than those covered by norm-referenced tests. Tyler's arguments were well received throughout American education, and by the mid-1940s Tyler had set the stage for exerting a heavy influence on how educators and other program evaluators viewed evaluation for the next twenty-five years.

The Age of Innocence: 1946 to 1957

In the ensuing years, Tyler's recommendations were more discussed than applied. Throughout American society, the late 1940s and 1950s were a time to forget the war, leave the depression behind, build and expand capabilities, acquire resources, and engineer and enjoy a good life. We might have called this era the Period of Expansion, except that there was also widespread complacence in regard to serious societal problems. We therefore think this time is better referred to as the Age of Innocence, or even as the Age of Social Apathy.

More to the point of educational evaluation, there was expansion of educational offerings, personnel, and facilities. New buildings were erected. New kinds of educational institutions, such as community colleges, emerged. Small school districts consolidated with others to provide the wide range of educational services that were common in larger school systems: mental and physical health services, guidance, food services, music instruction, expanded sports programs, business and technical education, and community education. Enrollment in teacher education programs ballooned, and college enrollment generally increased dramatically.

This general scene in society and education was reflected in educational evaluation. Although there was great expansion of education, society had no particular interest in holding educators accountable, identifying and addressing the needs of the underprivileged, or identifying and solving problems in the U.S. education system. Although educators wrote about evaluation and collected considerable data, they seem not to have related these efforts to attempts to improve educational services. This lack of a mission carried over into the development of the technical aspects of evaluation as well. There was considerable expansion of tools and strategies for applying the various approaches to evaluation: testing, comparative experimentation, operationalizing objectives, and comparing outcomes and objectives. As a consequence, educators were provided with new tests and test scoring services, algorithms for writing behavioral objectives, taxonomies of objectives, new experimental designs, and new statistical procedures for analyzing educational data. But these contributions were not derived from any analysis of what information was needed to assess and improve education, and they were not an outgrowth of school-based experience.

During this period, educational evaluations were, as they had been previously, primarily the purview of local school districts. Schools could do evaluation or not, depending on local interest and expertise. Federal and state agencies had not yet become deeply involved in the evaluation of programs. Funds for evaluations came from local coffers, foundations, or professional organizations. This lack of external pressure and dearth of support for evaluations at all levels of education would end with the arrival of the next period in the history of evaluation.

The Age of Realism: 1958 to 1972

The Age of Innocence in evaluation came to an abrupt end in the late 1950s and early 1960s with the call for evaluations of large-scale curriculum development projects funded by federal monies. Educators would find during this period that they no longer could do or not do evaluations as they pleased, and that further development of evaluation methodologies would have to be grounded in concern for accountability, usability, and relevance. Their rude awakening during this period would mark the end of an era of complacency and help launch profound changes, guided by the public interest and dependent on taxpayer monies for support, which would help evaluation expand as an industry and into a profession.

The federal government responded to the Russian launch of Sputnik I in 1957 by enacting the National Defense Education Act of 1958. Among other things, this act provided for new educational programs in mathematics, science, and foreign languages and expanded counseling and guidance services and testing programs in school districts. A number of new national curriculum development projects, especially in science and mathematics, were established. Eventually funds were allocated to evaluate these programs.

Four approaches to evaluation were represented in the evaluations done during this period. First, the Tylerian approach was used to help define objectives for the new curricula and to assess the degree to which the objectives were later realized. Second, new nationally standardized tests were developed to better reflect the objectives and content of the new curricula and to begin monitoring the educational progress of the nation's youth (L. V. Jones, 2003). Third, the professional judgment approach typically engaged experts to rate proposals and make periodic site visits to check on the efforts of contractors. Finally, many evaluators studied curriculum development efforts through the use of controlled field experiments.

In the early 1960s some leaders in educational evaluation realized that their work and their results were not particularly helpful to curriculum developers or responsive to the questions about the programs being raised by those who wanted to assess their effectiveness. The "best and the brightest" of the educational evaluation community were involved in these efforts to evaluate the new curricula; they were adequately financed, and they carefully applied the technology that had been developed during the past decade or more. Despite all this, they began to recognize that their efforts were not succeeding.

This negative assessment was well reflected in a landmark article by the educational psychologist Lee Cronbach (1963). In looking at the evaluation efforts of the recent past, he sharply criticized the guiding conceptualizations of evaluation for their lack of relevance and utility and advised evaluators to turn away from their penchant for evaluations based on comparisons of the norm-referenced test scores of experimental and control groups. Cronbach counseled evaluators to reconceptualize evaluation not in terms of a horse race between competing programs, but instead as a process of gathering and reporting information that could help guide curriculum development. Cronbach argued that analysis and reporting of test item scores would be likely to prove more useful to teachers than the reporting of average total scores. Initially, Cronbach's counsel and recommendations went largely unnoticed except by a small circle of evaluation specialists. Nonetheless, his article was seminal, containing

hypotheses about new approaches to conceptualizing and conducting evaluations that were to be developed and tested within a few years.

The War on Poverty was launched in 1965. It was grounded in the previous pioneering work of Senator Hubert Humphrey and the charismatic leadership of President John F. Kennedy before his untimely death in 1963. President Lyndon Johnson subsequently picked up the reins and used his great political skill to get this landmark legislation passed. Its programs poured billions of dollars into reforms aimed at equalizing and upgrading opportunities for all U.S. citizens across a broad array of health, social, and educational services. The expanding economy enabled the federal government to finance these programs, and there was widespread support throughout the nation for developing what President Johnson termed the Great Society. Accompanying this massive effort to help those in need was a concern in some quarters that the investments might be wasted if appropriate accountability requirements were not imposed.

In response to this concern, Senator Robert Kennedy and some of his colleagues in Congress amended the Elementary and Secondary Education Act of 1965 to include specific evaluation requirements. As a result, Title I of that act (aimed at providing compensatory education to disadvantaged children) specifically required each school district receiving funds under this title to evaluate Title I projects annually using appropriate standardized test data and thereby to assess the extent to which the projects had achieved their objectives.

This requirement, with its specific reference to standardized test data and an assessment of congruence between outcomes and objectives, reflects the state of the art in educational evaluation at that time, which was based largely on the use of standardized educational achievement tests and superficially on Tyler's objectives-based approach. More important, the requirement forced educators to move their concern for educational evaluation from the realm of theory and supposition into the realm of practice and implementation. When school districts began to respond to the evaluation requirements of Title I, they quickly found that the existing concepts, tools, and strategies employed by their evaluators were largely inappropriate for the task.

Available standardized tests had been designed to rank-order students of average ability; they were of little use in diagnosing needs and assessing the gains of disadvantaged children whose educational development lagged far behind that of their middle-class peers. Furthermore, these tests were found to be relatively insensitive to differences between schools and programs, mainly because of their psychometric properties and content coverage. Instead of being measures of outcomes directly relating to a school or a particular program, these tests were at best indirect indicators of learning, measuring much the same traits as general ability tests (Kellaghan, Madaus, & Airasian, 1982).

The use of standardized tests entailed another problem, because it conflicted with the precepts of the Tylerian approach. Because Tyler recognized and encouraged differences in objectives from locale to locale, this model became difficult to adapt to nationwide standardized testing programs. To be commercially viable, these standardized testing programs had to overlook, to some extent, objectives stressed by particular locales in favor of objectives stressed in the majority of districts.

Also, the Tylerian rationale itself proved inadequate to the evaluation task. There was insufficient information about the needs and achievement levels of disadvantaged children to

guide teachers in developing meaningful behavioral objectives for this population of learners. In retrospect, the enormous investment school districts across the United States made in training and leading educators to write behavioral objectives was largely unsuccessful and a waste of much time and money. Typically educators learned how to meet the technical requirements of good behavioral objectives. However, these technically sound statements of objectives often proved to be of little practical use in that they did not reflect empirical assessments of the needs and problems of the students to be served. When the teachers actually met their students, they often found it prudent to set aside as irrelevant the objectives that had been so carefully prepared in advance of the project.

Attempts to isolate the effects of Title I projects through the use of experimental and control group designs also failed. Typically such studies showed "no significant differences" in achievement between treated Title I students and comparison groups. This approach was widely tried but was doomed not to succeed. Title I evaluators could not begin to meet the assumptions required by experimental designs. For example, they usually could not, in a timely manner, obtain valid measures; could not hold treatments constant during the study period; and legally could not randomly assign Title I (disadvantaged) students to control and experimental groups. When the finding of no results was reported, as was generally the case, there was little information on what the treatment was supposed to be and often no data on the degree to which it had in fact been implemented. Also, the emphasis on pre- and posttest scores diverted attention from consideration of the treatment or of treatment implementation. This hugely expensive experiment in testing the utility and feasibility of experimental design evaluations in the Title I program demonstrated rather decisively that this technique is not amenable to evaluating highly dynamic, field-based, generalized assistance programs, especially in the course of such programs' development.

As a result of growing disquiet concerning evaluation efforts and consistently negative findings, Phi Delta Kappa set up the National Study Committee on Evaluation (Stufflebeam et al., 1971). After surveying the scene, this committee concluded that educational evaluation was seized with a great illness and called for the development of new theories and methods of evaluation as well as for new training programs for evaluators. This committee's indictment of educational evaluation practice was consistent with a study of government-sponsored evaluations by Guba (1966) and an analysis of the Title I evaluation efforts by Stufflebeam (1966b).

At the same time, many new conceptualizations of evaluation began to emerge. Provus (1969), Hammond (1967), Eisner (1975), and Metfessel and Michael (1967) proposed reformulations of the Tylerian model. R. Glaser (1963), R. W. Tyler (1967), and Popham (1971) pointed to criterion-referenced testing as an alternative to norm-referenced testing. D. L. Cook (1966) called for the use of system analysis techniques to evaluate programs. Scriven (1967, 1974); Stufflebeam (1967); Stufflebeam et al. (1971); and Stake (1967) introduced new models for evaluation that departed radically from prior approaches. These conceptualizations stemmed from recognition of the need to evaluate goals, look at inputs, examine implementation and delivery of services, and measure intended as well as unintended program outcomes. Developers of these new approaches also emphasized the need to make (or collect) judgments about the merit and/or worth of the object being evaluated.

The late 1960s and early 1970s were vibrant with descriptions, discussions, and debates concerning how evaluation should be conceived. The chapters in Part Three deal in depth with the alternative approaches that began to take shape during this period. Lessons had been learned, often by uneasy experience.

The Age of Professionalism: 1973 to 2004

Beginning in about 1973, the field of evaluation began to crystallize and emerge as a distinct profession in its own right—related to, but quite distinct from, its forerunners of research and testing. The field of evaluation has advanced considerably as a profession, yet it is instructive to consider the development in the Age of Professionalism in the context of the field in the previous period.

In the late 1960s and early 1970s, evaluators faced an identity crisis. They were uncertain of their role—whether they should be researchers, testers, reformers, administrators, teachers, consultants, or philosophers. What special qualifications, if any, they should possess was unclear. There were no professional organizations dedicated to evaluation as a field, nor were there specialized journals through which evaluators could exchange information about their work. Essentially no literature about evaluation existed, except for unpublished papers that circulated through a small underground network of scholars. There was a paucity of pre-service and in-service training opportunities in evaluation. Articulated standards of good practice were confined to educational and psychological tests. The field of evaluation was amorphous and fragmented. Many evaluations had been conducted by untrained personnel or research methodologists who had tried unsuccessfully to fit their experimental methods to evaluations (Guba, 1966). Evaluation studies were fraught with confusion, anxiety, and animosity. Evaluation as a field had little stature and no political clout.

Against this backdrop, the progress made by evaluators in professionalizing their field beginning in the 1970s is quite remarkable. Many universities now offer at least one course in evaluation methodology (as distinct from research methodology). A few—including the University of Illinois, the University of California at Los Angeles, the University of Minnesota, the University of Virginia, Claremont Graduate University, and Western Michigan University—have developed graduate programs in evaluation (LaVelle & Donaldson, 2010). Even so, the Western Michigan University program is the world's only interdisciplinary doctoral program in evaluation (Coryn, Stufflebeam, et al., 2010).

Increasingly, the field has looked to metaevaluation (Scriven, 1975; Stufflebeam, 1978, 2001c) as a means of ensuring and checking the quality of evaluations. In 1981 the Joint Committee issued standards for judging evaluations of educational programs, projects, and materials and established a mechanism by which to review and revise the standards and assist evaluators in using them. This review process has worked effectively, leading the Joint Committee to produce the second edition of *The Program Evaluation Standards* in 1994 and the third edition in 2011.[4] Moreover, publication of the Joint Committee's first edition of *The Personnel Evaluation Standards* in 1988, followed by the second edition in 2009, signaled advancement in methods for assessing systems for evaluating personnel. In addition, the Joint Committee

released *The Student Evaluation Standards* in 2003. Several other sets of standards with relevance for evaluation also have been published, the most important being AEA's *Guiding Principles for Evaluators* (2004) and the U.S. Government Accountability Office's *Government Auditing Standards* (U.S. General Accounting Office, 2002; U.S. Government Accountability Office, 2003, 2007). Many new techniques and methodological approaches have been introduced for evaluating programs, as described in Part Four of this book. The most comprehensive treatment of the state of the art in educational evaluation so far is the *International Handbook of Educational Evaluation* (Kellaghan & Stufflebeam, 2003).

The Age of Global and Multidisciplinary Expansion: 2005 to the Present

When the first edition of this book was being completed in 2006 (Stufflebeam & Shinkfield, 2007), it was realized that the evaluation field had already entered a new, recognizable age. Here, we label it the Age of Global and Multidisciplinary Expansion and arbitrarily have set its beginning as about 2005. As noted earlier, there are now over fifty professional evaluation societies in countries throughout the world (for example, France, Norway, Sri Lanka), many of which were established during this period. Moreover, the growing evaluation profession encompasses a wide range of disciplines and evaluators from various disciplinary perspectives and backgrounds who increasingly are exchanging information, studying in interdisciplinary degree programs, working on evaluation projects together, publishing together, and meeting together in broadly focused evaluation conventions and meetings. The last type of interaction is reflected in the Evaluation Conclave in southern Asia, which held its first conference in New Delhi, India, in 2010. As previously alluded to, the Canadian Evaluation Society initiated its Credentialed Evaluator (CE) designation in 2010, with designation meaning that the holder has provided adequate evidence of having obtained the education and experience required to be considered a competent evaluator.

Our own experience at Western Michigan University and elsewhere is applicable here, because in 2003 we established the first Interdisciplinary PhD in Evaluation program (Coryn, Stufflebeam, et al., 2010). This program's instructors, advisers, and students have backgrounds in such diverse disciplines as nursing, substance abuse treatment, sociology, social work, business, community development, economics, education, engineering, psychology, chemistry, public administration, statistics, and political science. The evaluation-related learning experiences of both students and faculty members are greatly enhanced by students' conducting fieldwork projects together. Also in 2004, under Scriven's leadership, the IDPE program established the open-access, online *Journal of MultiDisciplinary Evaluation,* originally modeled after the *Harvard Law Review*. This journal has been widely subscribed to across disciplines and internationally. Clearly, the evaluation profession is becoming increasingly pervasive in disciplines and nations across the world.

During this period, many evaluation sponsors in the United States and elsewhere have returned to requiring so-called evidence-based evaluation methods. Generally, these are patterned after the evidence-based practice model in medicine (that is, randomized controlled trials). This approach is now often required for evaluating both independent and federally sponsored initiatives charged with identifying effective interventions (U.S. Government

Accountability Office, 2009). The reemerging federal requirements for applying experimental design mirror similar requirements that were previously advocated by Campbell and others in the 1960s (Campbell & Stanley, 1966). Also during this period, many alternative evaluation models and approaches, developed and prescribed in earlier periods in the history of evaluation, have gained greater prominence, legitimacy, and application. Among these are transformative evaluation, appreciative inquiry, participatory evaluation, empowerment evaluation, and theory-driven evaluation (Coryn, 2009).

It also is notable that during this period many long-standing disagreements among members of the evaluation community have resurfaced. In particular, disagreements about appropriate methods for inferring cause-and-effect relationships between programs and their outcomes as well as the persistent quantitative-qualitative debate (T. D. Cook, Scriven, Coryn, & Evergreen, 2010; Donaldson & Christie, 2005; Donaldson, Christie, & Mark, 2009), which for a short period diminished, have again intensified in the field. Relatedly, many international organizations, such as the World Bank and similar entities, which historically have relied on experimental and econometric methods of evaluation, have slowly begun a shift toward participatory evaluation, theory-driven evaluation, self-evaluation, and other alternative models and approaches for evaluating their humanitarian efforts. Also, the U.S. Government Accountability Office (2007) intensified its position that audits and other evaluations of federal programs must meet requirements for independence and objectivity.

Summary

In this chapter we have made the following points:

- Societies and their institutions require formal, systematic evaluations, as distinguished from everyday, informal evaluations (which are inevitable and also often lacking in reliability).
- The definitions of formal evaluation we provided are keyed to values (merit, worth, and probity); assessment criteria; and needed interface/communication and technical tasks.
- Key criteria for judging programs include quality, accomplishments, side effects, responsiveness to assessed needs, cost-effectiveness, probity, safety, sustainability, transportability, and others.
- The main functions of evaluation are formative and summative.
- Basically, noncomparative approaches are appropriate for evaluating programs under development, whereas comparative approaches often are needed to evaluate completed programs.
- Evaluation is a profession that serves all other professions and draws from the full range of disciplines.
- Professionalism requires one to obtain and use evaluation to increase competence and improve services.
- The professionalization of evaluation over time has been tied closely to the field of education and has occurred across the Pre-Tylerian Period, the Tylerian Age, the Age of

Innocence, the Age of Realism, the Age of Professionalization, and the Age of Global and Multidisciplinary Expansion.

- Evaluations themselves must be assessed against the standards of the evaluation field—for example, those developed by the Joint Committee and the U.S. Government Accountability Office.

REVIEW QUESTIONS

1. List, contrast, and discuss the benefits and limitations of formal evaluations.
2. Explain and give examples of evaluation's symbiotic relationships with other fields.
3. Cite what you see as the pros and cons of defining evaluation as a process of comparing outcomes to objectives and, conversely, the pros and cons of defining evaluation as the systematic assessment of merit and worth.
4. Summarize this chapter's stated rationale for employing values clarification in program evaluations; list and explain key issues in clarifying the values held by a program's stakeholders; and then list steps that you see as potentially effective for clarifying stakeholder values and applying them to reach evaluative conclusions.
5. Describe an example of how members of U.S. society were put at risk or harmed due to the failure of responsible parties to heed and act on the findings of an evaluation.
6. Cite some reasons why evaluators should search for side effects.
7. Suppose you want to increase your competence to conduct program evaluations. List and give examples of the main categories of skills you would seek to acquire, and discuss how you believe you could best obtain these skills.
8. Define what is meant by the terms *merit* and *worth*. Then, from your experience, write an example of a program or other entity that possessed merit but not worth. Describe how merit and worth were assessed. Explain why assessments of worth are dependent on context.
9. Give examples of cases that require comparative evaluations. Give examples of other cases that require only noncomparative evaluations.
10. Compare and contrast the terms *formative evaluation* and *summative evaluation*. Give an example of each of these evaluation roles.

Group Exercises

This section is designed to support group discussion of key issues addressed in this chapter. Each exercise summarizes a particular case, then provides instructions for the group's analysis of and response to the case. After your group's members have read an exercise, engage in discussion to arrive at your group's response to the particular assignment.

Exercise 1

The head of a large state government department has found himself under political pressure to commission an evaluation of each of the four divisions of his department. None of these divisions has ever been evaluated except in the most cursory fashion, and then only sporadically. What is evident to stakeholders (the public) is that services of all four departments are costly but inadequate, and that the poor quality of delivery is causing growing frustration. Realistic financial provisions and timelines have been made available for this major evaluation, according to the head of the department. Suppose your group has been selected to conduct the evaluation. Outline the important early decisions you would need to make about key aspects of the evaluation; the kinds of initial understandings you would need to reach with the head of the department and division heads; and the kinds of assurances you would seek and give so that a successful evaluation can eventuate.

Exercise 2

A superintendent of a small school district is beset with problems relating to the introduction of a new state-mandated science program for grades 7 through 9. She has heard of both formative and summative evaluation processes, but has little grasp of their functions and possible benefits if applied to the new science program. Your services are engaged to give the superintendent a thorough understanding of what constitutes formative and summative evaluation. Outline the relevance of either form of evaluation to the superintendent's problems, suggest a circumstance under which formative evaluation might lead to summative evaluation, and state the kind of cooperation an evaluation team would find essential to completing a successful evaluation. What advice do you give the superintendent?

Exercise 3

As a group, identify two studies: one that meets the requirements of a sound research investigation but not those of a summative evaluation, and one that meets the requirements of a sound summative evaluation. Then construct a matrix that shows the main distinctions and similarities between the two types of studies. Subsequently, discuss whether the distinctions your group identified are real and important.

Exercise 4

As a group, list points for use in explaining the essential differences between informal and formal evaluation, and also between formative and summative evaluation.

Notes

1. The Joint Committee is a standing committee that was established in 1975. Its approximately eighteen members have been appointed by about fifteen professional societies in the United States and Canada that are concerned with improving evaluations in education. The committee's charge is to develop

standards for educational evaluations. So far, it has created standards for evaluations of educational programs, personnel, and students. This book's first author was the committee's founding chair.

2. Although the Joint Committee expanded its definition of evaluation in the third edition of *The Program Evaluation Standards* (2011, p. xxv), we prefer the 1994 definition and refer to it throughout the chapter.
3. This section on the history of educational evaluation is largely based on a previous account by Stufflebeam, Madaus, and Kellaghan (2000), which included Madaus's incisive analysis of the early history of educational testing and evaluation.
4. The initial 1981 edition was titled *Standards for Evaluations of Educational Programs, Projects, and Materials*. For convenience, however, throughout this book we refer to all three editions as *The Program Evaluation Standards* and by the year of publication.

Suggested Supplemental Readings

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2010). *Program evaluation: Alternative approaches and practical guidelines* (4th ed.). Upper Saddle River, NJ: Pearson.

Fournier, D. M. (1995). Establishing evaluative conclusions: A distinction between general and working logic. In D. M. Fournier (Ed.), *Reasoning in evaluation: Inferential links and leaps* (pp. 15–32). New Directions for Evaluation, no. 68. San Francisco, CA: Jossey-Bass.

Kellaghan, T., & Stufflebeam, D. L. (Eds.). (2003). *International handbook of educational evaluation*. Norwell, MA: Kluwer.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

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

Scriven, M. (1993). *Hard-won lessons in program evaluation*. New Directions for Program Evaluation, no. 58. San Francisco, CA: Jossey-Bass.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Thousand Oaks, CA: Sage.

Shadish, W. R., & Luellen, J. K. (2005). History of evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 183–186). Thousand Oaks, CA: Sage.

CHAPTER 2

EVALUATION THEORY

This book's central purpose is to present an organized summary of the major contemporary approaches to program evaluation, followed by guidelines for applying the approaches. As background for these tasks, this chapter addresses the fundamental issue of evaluation theory. We believe that in the ongoing development of the evaluation field, evaluation approaches should be assessed and improved toward the goal of meeting the requirements of a sound theory (N. L. Smith, 1993). Given that goal, there is a need for evaluation scholars and evaluation graduate students to contribute to developing and validating sound evaluation theory. In this chapter we present our case for the needed evaluation theory development, and we suggest concepts, definitions, criteria, and related ideas for assisting with that work. This chapter's discussion is framed generally in terms of overall evaluation theory but concentrates more specifically on program evaluation theory. Later in the book, we refer back to this chapter to identify dimensions for use in characterizing and evaluating a range of different program evaluation approaches.

General Features of Evaluation Theories

It is possible to distinguish between general and specific theories of evaluation. A general metatheory of evaluation would characterize the nature of evaluations, regardless of subject matter, time, and space (Fournier, 1995; Scriven, 1991; Shadish, Cook, & Leviton, 1991). Such a general theory would cover a wide range of evaluations; denote their modal characteristics, including the logic and processes of evaluative discourse; and describe in general how evaluations should be assessed and justified. Specific theories of

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **The role of theory in advancing the evaluation discipline**
- **The contrast between an evaluation theory and an evaluation approach or model**
- **Program evaluation theory as only one of many types of evaluation theories**
- **Definitions of sound evaluation theory and its components**
- **Alternative sets of criteria for judging an evaluation theory**
- **The rationale for viewing theory development as a creative process that defies prescriptive approaches**
- **An assessment of progress toward developing validated theories of program evaluation**
- **The importance and difficulties of considering context in theories of program evaluation**
- **Illustrative hypotheses to be tested in the course of developing a theory of program evaluation**
- **The potential uses of grounded theory, professional evaluation standards, and metaevaluations in testing and improving evaluation theories**

evaluation would have many of the same characteristics as a general metatheory, but they would be delimited to account for evaluations that are restricted to particular types of evaluands, substantive areas, locations, or time periods.

It is important to note that theories focused on program evaluations constitute only one part of the general area of evaluation theory. Other parts include, for example, theories concerned with evaluations of personnel, commercial products and services, organizations, manufacturing, governance, policies, and even evaluation theories. Further divisions of evaluation theory are possible when one considers a range of disciplines. One could generate theories of different types of program evaluations—for example, evaluations of educational, social, economic, environmental, missionary, foreign aid, law enforcement, national defense, fundraising, engineering, and business programs. It would therefore be possible to consider general and specific theories of evaluation for the evaluation field as a whole or for any of its subareas.

In general, the evaluation profession spans evaluations of all sectors of society. All sectors employ programs, that is, interrelated sets of goal-directed activities. It is in both public and private interests to evaluate programs to enhance prospects for success and help ensure accountability to sponsors and constituents. Program evaluation crosses many disciplines and fields of service, is one of the most fully developed and important parts of the broader evaluation field, and is worthy of close study.

Providing a comprehensive view of the development of program evaluation theory is not as challenging as analyzing theory in more mature fields, such as economics, physics, jurisprudence, and psychotherapy. Although evaluation theorists have advanced creative and influential models and approaches for conducting program evaluations, these constructions have not been accompanied by a substantial amount of related empirical research. Consequently, no vast body of evidence exists on the functioning of different evaluation approaches. This is so partly because the young evaluation field has been engaged in theory development during a much shorter period of time than have more mature professions. Also, program evaluation scholars tend to be pragmatists (Scriven, 1998). Rather than trying to understand the relationships among variables in program evaluations as they play out in the real world, these scholars have concentrated on providing evaluators with new evaluation approaches and tools designed to improve evaluation practice.

Basically, program evaluation scholars have sought to develop approaches that assist evaluators in designing and carrying out useful, defensible program evaluations. For the most part, the program evaluation scholars and other evaluation researchers have not systematically generated and tested propositions from their conceptualizations of program evaluation nor used such findings to improve those conceptualizations (N. L. Smith, 1993). Thus, the program evaluation field lacks a sufficient body of research and steadily improving theories flowing from an ongoing process of rigorous, empirically grounded theory development. Nevertheless, the creative, influential constructions of a range of conceptual leaders in evaluation are intriguing, have been influential, and are worthy of scrutiny. They might aptly be termed "pre-theories" or "emergent theories." We do most of our analysis of these constructions in Parts Two and Three of this book when we describe and examine various evaluation approaches.

Theory's Role in Developing the Program Evaluation Field

Program evaluation, an important field of professional practice, is in its early stages of development. Only recently has the broader society (especially in the United States and Canada) begun to recognize evaluators as members of an identifiable, creditable profession. Like all other professionals, members of this emergent field need to study and continually improve their services. Their aim should be to produce a science of program evaluation—one that not only is grounded in ongoing conceptualization and rigorous testing of theory-based propositions but also continually improves. Theory often has been cited as one of the most useful of all things because it informs practice (Lewin, 1952). Reciprocally, feedback from practice is needed to validate and strengthen theories.

Use of perhaps overly narrow evaluation theories can lead to certain negative as well as positive consequences. On the one hand, for example, the theory that defined evaluation as the process of determining whether specific objectives have been achieved misled evaluators for decades as they focused only on intended outcomes, not on the crucially important side effects seen in many programs or on context and process. On the other hand, when Ralph W. Tyler introduced his famous objectives-based approach to evaluation in the 1930s, he provided a valuable service by giving educators a framework for systematically determining whether educational innovations were achieving their stated goals (Madaus & Stufflebeam, 1988). We see in this paragraph's apparent contradiction about the value of objectives-based evaluation that theories have positive and negative influences, may have differential utilities reflecting the conditions and needs in different eras, should be subjected to ongoing examination and reformulation, and should be recommended for use only if appropriate caveats are observed.

Another example of the ability of theories to mislead is seen in the shackling position, held by many influential parties especially in the 1960s and 1970s, that most program evaluations should employ strict experimental research methods, especially as applied in laboratories. Historically, this proposition, often backed by government mandates and funding restrictions, had a crippling, wasteful influence on the practice of program evaluations. Over the past three decades, evaluation leaders have made compelling arguments against such sweeping requirements (for example, Guba, 1969; Schwandt, 2004; Stake, 1975b, 1988), and many proponents of the approach have counseled that it be used only when circumstances make it a viable option. Here we see that program evaluations are embedded in societal dynamics and highly subject to political forces. It follows that sound program evaluation theories should account for cultural context and relevant political dynamics.

In fact, program evaluation theorists have sometimes played dissident roles in the development of the program evaluation discipline. In their own times, some of the field's creative leaders have been rebels: rebels against the philosophy of evaluation as a value-free science, against the philosophy of positivism, against the designation of laboratory experimental methods as the gold standard of field studies, against the penchant for pursuing scientific adequacy in evaluations to the exclusion of utility, and against the dominance of standardized testing in evaluating educational programs.

Fortunately, the program evaluation field has been blessed with a number of creative theorists who not only attacked what they saw as debilitating traditions in evaluation but also advanced alternative conceptualizations. Among these are Robert Stake (1976), who called for responsive rather than preordinate evaluations; Michael Scriven (1973), who advocated goal-free evaluation as a countermeasure to the narrowness of goals-based evaluation; Lee Cronbach (1982), who advocated contingency-based evaluation, with an emphasis on generalizability; and Egon Guba (1978; Guba & Lincoln, 1989), who proposed a naturalistic approach as opposed to the still in-vogue experimental design approach (also see Donaldson, Christie, & Mark, 2009). These theorists have contributed to profession-wide dialogue on the meaning of evaluation and its appropriate uses in real-world settings (also see T. D. Cook, Scriven, Coryn, & Evergreen, 2010; Davidson, 2007; Donaldson & Christie, 2005).

Functional and Pragmatic Bases of Extant Program Evaluation Theory

It is important to note that these and other evaluation theorists largely drew their ideas from practical experience (also see Chelimsky, 1998). They remained close to field experience data, drew their creative reconstructions based on these data, and maintained a functional orientation. The differences among the constructions of the program evaluation theorists no doubt stem in part not only from their different worldviews and philosophies but also from their different evaluation experiences (also see Alkin, 2004). The historical link between theoretical contributions and practical application has remained evident throughout the development of the program evaluation field. Theorists' conduct of evaluations in widely differing settings no doubt heavily influenced their different constructions of program evaluation. This fact argues strongly that program evaluation theories should direct evaluators to take explicit account of the contextual conditions surrounding their evaluations.

The fact that leading evaluation theorists have never totally embedded their work in the mainstream of empirical research methodology has had several important implications for the program evaluation field. On the one hand, it has tended to free program evaluation theory from the grip of conventional modes of thought and preconceptions concerning the conduct of field studies. By being relatively uninvolved in the ongoing institution of formal scientific inquiry, program evaluation theorists have more easily been able to question or reject assumptions that were patently accepted by traditional laboratory researchers and to make creative contributions. On the other hand, this lack of involvement has also freed them from some of the discipline of and responsibility for reasonably systematic and organized formulation and testing of hypotheses, which are the heritage of the well-socialized laboratory researcher. One of our main reasons for writing this chapter is that although we believe evaluation theorists need to continue producing creative, even rebellious conceptualizations of evaluation approaches, we feel that they also must proceed to derive and formally test theoretical propositions about the proper, effective conduct of evaluation, and then reformulate their theories pursuant to the obtained empirical evidence.

A Word About Research Related to Program Evaluation Theory

Several recent investigations have looked at the extent to which evaluation practitioners have applied theorists' recommended program evaluation approaches in practice (for example, Birckmayer & Weiss, 2000; Christie, 2003; Coryn, Noakes, Westine, & Schröter, 2011; Cullen, Coryn, & Rugh, 2011; R. L. Miller & Campbell, 2006). Christie (2003), for example, addressed this issue by developing a framework to compare reported practices of eight evaluation theorists with reported practices of a group of practitioners who had been evaluating California's Healthy Start Program. Essentially she wanted to learn something about whether evaluation practice mirrors evaluation theory. In response to her survey, a small percentage of practitioners reported using any of the eight theorists' approaches to evaluation. Datta (2003) reported that this finding was consistent with those from a few other studies, which found that practicing evaluators pay little attention to recommended theoretical approaches to evaluation. Datta warned, however, that the existing evidence on this issue was thin and had noteworthy limitations, especially a lack of generalizability.

Whether or not Christie's findings (2003) hold under further investigation, we think they raise additional fundamental questions. If it is generally true that evaluators do not apply evaluation theory, then it is important to ask why they do not. Perhaps the approaches are not sufficiently articulated for practical use, or the practitioners are not competent to carry them out, or the approaches lack convincing evidence that their use produces the needed evaluation results. Or it is possible that evaluators become so accustomed to using only one approach or very few that they become complacent and indifferent to the value of alternative models. Such considerations as examining, using, and contributing to theory may not feature prominently, if at all.

We think the first three explanations are plausible and should be studied. The third explanation has particular salience for this chapter. It seems understandable that practitioners, however well trained in the discipline of evaluation, would be unlikely to apply, and to continue to apply, any evaluation theory unless research had shown that when the theory is applied correctly, its use produces sound evaluation results.

An analogy from medicine may help clarify this point. Consider the theory of heart transplant surgery first developed at the Mayo Clinic. It would have been unthinkable to advise the wide body of heart surgeons to transplant hearts using this new approach before Michael DeBakey and other prominent heart surgeons had thoroughly tested the procedure and shown it to succeed. Also, dissemination of this practice should have been restricted to surgeons who were properly trained and certified. Indeed, a finding from a study that surgeons in general were not employing the new heart transplant procedure would have brought welcome relief.

Although applications of House's deliberative democratic approach (House & Howe, 2000a, 2000b, 2000c) or Patton's utilization-focused evaluation (1997, 2008)—discussed in later chapters—do not portend possible dire outcomes similar to those of heart transplant surgery, the same principles apply. Theoretical approaches in any profession should be carefully researched and validated prior to advocating their widespread use, and those who are to apply

the approaches should be specifically trained and certified as competent in their correct application. These principles are hallmarks of any mature profession. From Christie's report (2003), it seems clear that evaluation participants in her sample were not thoroughly trained in the theoretical approaches being researched and maybe not in the logic and methodology of program evaluation in general. Quite possibly the findings from her study are more indicative of the primitive state of training and certification for evaluators and dissemination of new evaluation approaches than of the adequacies, inadequacies, and applicability of theoretical approaches to program evaluation.

Nevertheless, we acknowledge that the practice of evaluation is and should be pervasive. Informal or amateur evaluation is important, and the evaluation profession should offer conceptual, technical, and training assistance to service providers who have to conduct many of their own evaluations. To undergird this assistance, the evaluation field should develop, disseminate, and provide training in validated theories of evaluation.

Program Evaluation Theory Defined

Most of this book deals with program evaluation models or approaches and not the more advanced notion of program evaluation theories. We use the term *program evaluation model* to refer to an evaluation theorist's idealized conceptualization for conducting program evaluations. The experience born of trial and error, pragmatism, field practice, an ability to develop concepts creatively, and other factors may have contributed to theory underpinning a particular program evaluation model. Whatever the causes, members of the evaluation field have made substantial progress in developing program evaluation models, and these are valuable. Like theories, evaluation models and approaches also need careful scrutiny and testing. Parts Two and Three of this book are devoted to a critical review and analysis of the major program evaluation approaches (including those referred to as "models"). Although some writers would characterize those approaches as prescriptive theories, we have reserved the term *theory* for creatively developed yet more rigorously tested conceptualizations of program evaluation. We value the contributions of the evaluation model developers but believe the program evaluation field should seek a higher standard when determining what constitutes a theory of program evaluation. We have thus set more demanding requirements for theories than for evaluation models and approaches.

We find the following definition of a program evaluation theory to be useful (but not sufficient) for considering the scope and rigor required by sound theories of program evaluation. A program evaluation theory is a coherent set of conceptual, hypothetical, pragmatic, and ethical principles forming a general framework to guide the study and practice of program evaluation. This definition is useful for identifying features by which to classify and examine different theories. Although it does not explicitly identify criteria for evaluating a theory, we observe that such criteria become necessary at some stage. According to this definition, a sound program evaluation theory has six main features: overall coherence, core concepts, tested hypotheses concerning how evaluation procedures produce desired outcomes, workable procedures, ethical requirements, and a general framework for guiding program evaluation practice and conducting

research on program evaluation. Effectively addressing this definition's requirements is a worthy goal for developers of program evaluation theories but one that is elusive, far from achievement, and lacking specific criteria for judging theories. We acknowledge that some evaluation approaches rate well on coherence, core concepts, workable procedures, integrity, and guidance for research and practice. All of them, however, fall short in producing principles based on empirical research. We will comment briefly on the definition's requirements for conceptual, hypothetical, pragmatic, and ethical principles as they relate to the general field of program evaluation.

Conceptual Principles

The conceptual nature of program evaluation is evident in the evaluation literature. It contains a wide range of well-developed concepts, such as formative and summative evaluation (Scriven, 1967); constructivist and responsive evaluation (Guba & Lincoln, 1989; Stake, 1975a, 1975b, 2004a, 2004b); context, input, process, and product (CIPP) evaluation (Stufflebeam, 1967, 1971a, 2003a); utilization-focused evaluation (Patton, 1997, 2008); participatory evaluation (Cousins & Earl, 1992; Cousins & Whitmore, 1998); utility, feasibility, propriety, and accuracy standards for evaluations (Joint Committee on Standards for Educational Evaluation, 1994) as well as the new evaluation accountability standards introduced in the third edition of the Joint Committee's *The Program Evaluation Standards* (2011); and, related to the evaluation accountability standards, metaevaluation (Scriven, 1969b; Stufflebeam, 1978, 2001c). A useful set of definitions of such concepts appears in Scriven's *Evaluation Thesaurus* (1991) and in this book's glossary.

Hypothetical Principles

Work in developing research-based principles for conducting program evaluations has been virtually nonexistent. Research efforts to state and confirm hypotheses about what works in program evaluations and under what conditions have been lacking. This is a fertile area for doctoral dissertations and funded research on evaluation. Such hypothesis-testing research should take explicit account of the environmental circumstances surrounding the subject program evaluations and their guiding models and approaches. Later in this chapter we cite some illustrative general hypotheses about evaluation practices that we found in the evaluation literature and elsewhere.

Pragmatic Principles

Pragmatic principles denote ways of conducting evaluations that have been shown to work well in evaluation practice. Many valuable evaluation procedures and rules of thumb are available. These have grown from a vast amount of evaluation experience and are evident in such writings as the guidelines contained in *The Program Evaluation Standards* (Joint Committee, 1981, 1994, 2011) and the Fitzpatrick, Sanders, and Worthen (2011) *Program Evaluation: Alternative Approaches and Practical Guidelines* textbook, as well as in Part Four of this book. Procedural

recommendations derived from program evaluation practices are worthy of examination and validation by empirical research.

Ethical Principles

There has been progress in defining ethical principles for program evaluations. This is seen in *Guiding Principles for Evaluators,* published by the American Evaluation Association (AEA; 2004); the "Propriety" section of *The Program Evaluation Standards* (Joint Committee, 1994, 2011); and the writings of scholars (for example, Morris, 2003, 2008, 2011).

Application of our proposed definition of a program evaluation theory is not straightforward. Hypothetical principles cover as wide a range of activities as exist under the general rubric of program evaluation. The development of these into definable constructs, based on research, is nonexistent. We cite one such irksome area (of an endless number—what we cannot nail down is always irksome!). The fine line between intuition and fact has often been discussed and debated by evaluators. Those who give credibility to intuition support the importance of judgments by the evaluator and program stakeholders in program assessment. They contend that experience gives wisdom and insights that have very real value in making judgments and reaching conclusions. By contrast, those who adopt a strict empiricist approach insist on facts; they want to know what really is happening and have little interest in nonprofessional judgments of the occurrences. To complicate this theoretical issue, an intermediate group happily accedes to both views, depending on pertaining conditions. Is the intuitive adherent right? Is the empiricist right? Is there, in reality, a sharp dichotomy, or can there be a rational and accommodating fusion of the two points of view? There is as yet no best answer to these theoretical questions. This is one example where the application of high-quality, systematic research would shed light on unanswered questions. The development of program evaluation, including the theoretical constructs underlying the kinds of approaches we cover in this book, should rely increasingly on strong research.

Criteria for Judging Program Evaluation Theories

Beyond meeting the requirements of the preceding definition of a program evaluation theory, an evaluation theory needs to meet certain well-established criteria of a sound theory. Such criteria are seen in various definitions of theories in the professional literature. Scriven (1991) defined theories as "general accounts of a field of phenomena, generating at least explanations and sometimes also predictions and generalizations" (p. 360). Following from this definition (of general theories), leading criteria for evaluating a program evaluation theory are that it be useful in efficiently generating verifiable predictions or propositions concerning evaluative acts and consequences and that it provide reliable, valid, actionable direction for ethically conducting effective program evaluations. More specific additional criteria for evaluating the utility of a program evaluation theory, frequently referenced in writings on theory, include clarity and comprehensiveness of assumptions, parsimony, resilience, robustness, generalizability, and heuristic power.

General Criteria for Evaluating Evaluation Theories

General criteria for judging program evaluation theories are also evident in *The Program Evaluation Standards* (Joint Committee, 1994, 2011) and AEA's *Guiding Principles for Evaluators* (2004). The 2011 edition of *The Program Evaluation Standards* contains thirty specific standards that spell out requirements for program evaluations, including that they meet conditions of utility, feasibility, propriety, accuracy, and evaluation accountability. *Guiding Principles for Evaluators* requires that evaluators practice systematic inquiry, possess the needed competencies, meet conditions of integrity and honesty, steadfastly be respectful of people, and assume responsibility for serving the general and public welfare. It follows that sound theories of program evaluation should at least consider the professional standards and principles of the evaluation field. Exhibit 2.1 provides an organized list of some of the criteria for use in evaluating program evaluation theories.

Exhibit 2.1 GENERAL CRITERIA FOR EVALUATING PROGRAM EVALUATION THEORIES, ORGANIZED BY CATEGORY

Professionalizing Program Evaluation

Is the theory useful for . . .

- Generating and testing standards for program evaluations?
- Clarifying roles and goals of program evaluation?
- Developing needed tools and strategies for conducting evaluations?
- Providing structure for program evaluation curricula?

Research

Is the theory useful for . . .

- Generating and testing predictions or propositions concerning evaluative actions and consequences?
- Application to specific classes of program evaluation (the criterion of particularity) or a wide range of program evaluations (the criterion of generalizability)?
- Generating new ideas about evaluation (the criterion of heuristic power)?
- Drawing out lessons from evaluation practice to generate better theory?

Planning Evaluations

Is the theory useful for . . .

- Giving evaluators a structure for conceptualizing evaluation problems and approaches?
- Determining and stating comprehensive, clear assumptions for particular evaluations?
- Determining boundaries and taking account of context in particular program evaluations?

- Providing reliable, valid, actionable direction for ethically and systematically conducting effective program evaluations?

Staffing Evaluations

Is the theory useful for . . .

- Clarifying roles and responsibilities of evaluators?
- Determining the competencies and other characteristics evaluators need to conduct sound, effective evaluations?
- Determining the areas of needed cooperation and support from evaluation clients and stakeholders?

Guiding Evaluations

Is the theory useful for . . .

- Conducting evaluations that are parsimonious, efficient, resilient, robust, and effective?
- Promoting evaluations that clients and others can and do use?
- Promoting integrity, honesty, and respect for people involved in program evaluations?
- Responsibly serving the general and public welfare?

Shadish, Cook, and Leviton's Criteria for Theories of Program Evaluation

Prior to publication of the second and third editions of *The Program Evaluation Standards* (Joint Committee, 1994, 2011) and *Guiding Principles for Evaluators* (AEA, 2004), William Shadish, Thomas Cook, and Laura Leviton, in their influential book *Foundations of Program Evaluation: Theories of Practice* (1991), asserted that "judging the merits of evaluation theories requires specific description of the things that such theories ought to do and the issues they ought to address competently" (p. 36). More specifically, they declared that a comprehensive theory of program evaluation should fully address five general criteria: (1) social programming, (2) knowledge construction, (3) valuing, (4) use, and (5) practice. These five criteria are summarized and described in Table 2.1.

In their book, Shadish et al. (1991) applied these criteria to the prescriptive theories of evaluation developed and advocated by evaluation theorists Scriven, Donald Campbell, Carol Weiss, Joseph Wholey, Stake, Cronbach, and Peter Rossi. In doing so, they categorized the theories and theorists into three stages of historical and theoretical development. Stage I theories and theorists, beginning in the 1960s, provided the basic conceptual basis for valuing and knowledge construction, largely advocating rigorous scientific methods for doing so. Stage II theories and theorists, beginning in about the 1970s, recognized the complex nature of social interventions and programs as well as the limited use of evaluation findings in

Table 2.1 Shadish, Cook, and Leviton's Criteria for Theories of Evaluation

Criterion	Explanation
Social programming	The nature of social programs and their role in social problem solving
Knowledge construction	Acceptable knowledge claims about the object of evaluation, legitimate methods to produce knowledge claims, and assumptions about what kinds of knowledge are worth studying
Valuing	The role that values and the process of valuing play in evaluation, and how to construct value judgments
Use	How evaluative information is used, by whom, and for what purposes, and how to increase legitimate use
Practice	The things that evaluators do or should do in conducting evaluations

Source: Adapted from Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Thousand Oaks, CA: Sage.

policymaking. Theorists of this period therefore emphasized methods for getting evaluations used. Stage III theories and theorists, beginning in the 1980s, integrated prior theoretical work into developing comprehensive, contingency (if, then)–based approaches to evaluation in an effort to address what were perceived as deficiencies in prior theorists' theories of evaluation. Shadish et al.'s analysis of these theories and their respective theorists largely revealed that none adequately addressed all five criteria, though theories advocated by Cronbach and Rossi came closest. That being said, numerous other influential theorists (for example, Marvin Alkin, Eleanor Chelimsky, Elliot Eisner, Guba, Michael Patton, and Daniel Stufflebeam) who were writing about and practicing evaluation during the 1960s, 1970s, and 1980s were excluded from the analysis—although Shadish et al. recognized and acknowledged this exclusion. And much theoretical development has occurred in the twenty years since the book's publication, somewhat limiting its generalizability to contemporary theories of evaluation, though the authors' criteria may still be applicable.

Miller's Standards for Empirical Examinations of Evaluation Theories

More recently, R. L. Miller (2010) proposed a small set of standards for use in empirical investigations of the theory-practice relationship in evaluation. In doing so, she rightly noted:

> Although the benefits to evaluating how theories perform in practice seem obvious, there have been few attempts to examine theories in this way . . . Prior attempts to evaluate whether and how evaluation theories are put to practice suggest an emergent framework for empirically exploring how theory informs practice and whether particular theories of practice yield better evaluations. (p. 391)

R. L. Miller's standards (2010)—(1) operational specificity, (2) range of application, (3) feasibility in practice, (4) discernible impact, and (5) reproducibility—are briefly described in Table 2.2.

R. L. Miller's suggested standards (2010) for investigations of evaluation practice have important implications for what evaluators report in describing practice experiences and also

Table 2.2 Miller's Standards for Research on Evaluation

Criterion	Explanation
Operational specificity	Translates into clear guidance and sensitizing ideas for practitioners, and theoretical signature must be recognizable
Range of application	Identifies under what practice circumstances and in pursuit of what evaluative questions theory can be applied
Feasibility in practice	Ease or difficulty in applying theoretical prescriptions and sensitizing ideas in practice
Discernable impact	Identifies the impacts that are expected and desired and whether unintended effects occur
Reproducibility	Evaluation impacts that are observed can be reproduced over time, occasions, and evaluators

Source: Adapted from Miller, R. L. (2010). Developing standards for empirical examinations of evaluation theory. *American Journal of Evaluation, 31,* 390–399.

in regard to methodological criteria for selecting and assessing evaluations in which particular theories have allegedly been applied:

> Perhaps most obvious, cases must be described with adequate detail to allow others to study them. Details of importance would include clear statements of the evaluation setting, evaluation purpose, rationale for applying the theoretical approach, articulation of how the theory was enacted in the particular case, descriptions of all actors and their roles, a chronological event history of the evaluation, and information on what outcomes were expected to accrue from applying the approach and when and how these were substantiated. (p. 397)

Although R. L. Miller's standards (2010) have yet to be applied to investigations of evaluation practice and are themselves normative theoretical prescriptions, they probably do have value. Her standards for examinations of evaluation theories might serve as an insightful, informative complement to Part Two of this book, in which we assess a variety of evaluation approaches and models against conditions of utility, propriety, feasibility, accuracy, and evaluation accountability (Joint Committee, 2011).

Theory Development as a Creative Process Subject to Review and Critique by Users

Theory development, which is critically important to advancing program evaluation practice, is a creative, complex, difficult enterprise. Typically, sound theory development includes study of practice; taking account of context; making bounded, creative conceptualizations; operationalizing and applying the conceptualizations; rigorously studying applications; and revising conceptualizations. It is therefore properly conceived as an ongoing, cyclical, practice-linked, research-based, creative process that denotes the appropriate sphere of application.

Theory development basically is an exploratory, even arbitrary process. It is a matter of the free, creative choice of the theorist. Although we can outline the features of a sound theory, characterize a general cycle of ongoing theory development, and identify different writers' suggested criteria for evaluating theories, we will not lay out any one method for the development of program evaluation theories. We should not do so any more than we should

tell composers, poets, or artists how they must produce their creative contributions. In this sense, a theory of program evaluation is no more or less than an interrelated set of propositions about program evaluations created by the theorist. It stands on its own as a set of personalized predictions and propositions.

The theorist cannot, however, have her or his theory accepted and applied by just creating it and saying it is good. Like a musical composition, Broadway play, or painting, a program evaluation theory must pass muster with critics and the theorist's broader audience. Program evaluation theories are subject to evaluation by users, who will judge a theory of program evaluation to be useful or not based primarily on how efficiently and validly the theory generates verifiable predictions about the relationship between certain evaluation actions and evaluation outcomes as well as on propositions about how to carry out successful program evaluations that hold up in practice. It is up to the theorists to make and put forward their theories in whatever way they think is best. If they want their theories to count for something and be influential, then they will want to obtain rigorous research on the theories' utility and use the findings to improve the theories.

Status of Theory Development in the Program Evaluation Field

The relatively young evaluation profession has advanced substantially in conceptualizing the program evaluation enterprise, but it has far to go in developing overarching, validated theories to guide the study and practice of program evaluation (also see Shadish, 1994, 1998). The program evaluation literature's references to program evaluation theories are numerous, but these references are often pretentious. They usually denote as theories conceptual approaches or evaluation models that lack the comprehensiveness and validation required of sound theories.

Alkin's 2004 book, *Evaluation Roots: Tracing Theorists' Views and Influences,* is a case in point. It is valuable in its presentation of various conceptual approaches to program evaluation, but its labeling of these approaches as "theories" can be misleading. None of them meets the conditions for a fully developed, useful theory, nor can they be correctly identified as either descriptive or predictive. Clearly these conceptualizations do not provide validated predictions of the consequences of particular evaluation actions. Moreover, they do not hold up as general descriptions of how evaluations are actually carried out. Instead of characterizing how evaluations actually play out in practice, the conceptualizations referenced by Alkin mainly recount the particular authors' preferences concerning how evaluations should be done (that is, they are recommended approaches rather than descriptive or predictive theories). The conceptualizations presented by Alkin as theories lack evidence that evaluations are actually carried out in the ways described, which in itself is not a deficiency. However, the presented conceptualizations are more aptly labeled "prescriptive" rather than "descriptive" (or "predictive").

Nevertheless, we acknowledge that the program evaluation field has made substantial progress in conceptualizing approaches to program evaluation, and Alkin's book (2004) presents a valuable analysis of this progress. Also, evaluation theorists clearly have reflected on practical program evaluation experiences in conceptualizing their approaches and have sought

to make them useful. What we want readers to consider is that theory development in program evaluation still has far to go. We believe the weakest link in developing program evaluation theories so far is the lack of formulation and rigorous testing of hypotheses about the effects of applying different theoretical approaches in actual program evaluations. We hope sponsors of evaluation research will take note of the critical need to support studies to formulate and test hypotheses about what evaluative actions produce the most beneficial program evaluation outcomes under documented contextual circumstances.

Importance and Difficulties of Considering Context in Theories of Program Evaluation

Prospects for successful theory development in program evaluation are limited by difficulties inherent in predicting and generalizing in the social sciences. Any sound effort to develop a program evaluation theory needs to take into account the social, political, geographical, and temporal contexts of the program evaluations being studied. Such contexts vary widely in characteristics and influence from evaluation to evaluation. Moreover, the contexts for program evaluations typically are fluid, uncontrolled, and unpredictable. Without considering context, a theorist can hardly posit how a prescribed approach to evaluation will work or not work under any particular set of social, organizational, economic, and other conditions. Further, validated predictions, even in the physical sciences, may have a short half-life.

These difficulties concerning context give a view of the challenge of developing sound theories of program evaluation. We think they also underscore the point that development of program evaluation theories must be ongoing and that theories should regularly be assessed and updated. Furthermore, each program evaluation theory is best based on a wide range of program evaluations, both within particular types of contexts and across different types of contexts. In addition, we think program evaluation theories, if they are to be useful, should advise evaluators to assess and take account of each evaluation's unique context. This requirement is strongly made in the Explicit Program and Context Descriptions standard in *The Program Evaluation Standards* (Joint Committee, 2011) and the context evaluation component of Stufflebeam's CIPP model for evaluation (2003a).

Need for Multiple Theories of Program Evaluation

We think it noteworthy that evaluation's contradictory persuasions are not resolvable in any single, overall theory. Positivist, existentialist, constructivist, objectivist, and postmodern persuasions encompass irreconcilable philosophical differences. For example, objectivist approaches posit the existence of an underlying reality and charge evaluators to pursue this. But constructivist evaluators deny the existence of an underlying reality and call on evaluators to collect and report different, probably contradictory constructions of what is observed. And the existentialist gives particular emphasis to personal experience and responsibility in evaluations, a philosophy often exemplified in case studies. Such opposing conceptualizations of program evaluation can in their own terms be defensible, considering their different underlying precepts, assumptions, and experiences and the various ways they work out in practice.

Publications by Alkin (2004); Shadish et al. (1991); House (1983); Stufflebeam (2001b); Stufflebeam, Madaus, and Kellaghan (2000); Stufflebeam and Shinkfield (2007); and Kellaghan and Stufflebeam (2003) have acknowledged and presented fundamental differences among a wide range of individual conceptualizations of program evaluation. Although none of these conceptualizations meets the requirements of a fully validated theory, together they provide evaluators with a range of different approaches grounded in different philosophical persuasions and a wide range of experiences. We endorse efforts to evolve different, defensible conceptualizations of program evaluation into different validated theories for guiding the study and practice of program evaluation in particular types of settings and according to different philosophical approaches.

Hypotheses for Research on Program Evaluation

Efforts to develop program evaluation theory should include rigorous formulation and testing of hypotheses about what works in program evaluations, why, and under what conditions. A search for hypotheses drawn from the Evaluators' Institute courses, our own instructional and evaluation experiences, and the research literature revealed the following example hypotheses concerning different aspects of program evaluation:

Professional Standards and Principles for Program Evaluations

- Appropriate application of evaluation standards and principles enhances an evaluation's quality and contributes to the resolution of ethical problems (from Michael Morris's description of his 2004 Evaluators' Institute course; also see Joint Committee, 1981, 1994, 2011).
- Application of professional standards and principles can be used to solve ethical dilemmas in conducting evaluations (Morris, 2008, 2011).

Evaluation Approaches and Models

- Comprehensive evaluation approaches and models produce more credible, valid, and useful evaluations than do more limited approaches and models (Stufflebeam, 2001b; Stufflebeam & Shinkfield, 2007).
- Correct application of selected evaluation approaches and models produces their desired consequences (R. L. Miller, 2010; R. L. Miller & Campbell, 2006).

Involvement of Stakeholders

- Stakeholder involvement enhances use of evaluation findings (Alkin, Daillak, & White, 1979; Greene, 1988).
- Under certain conditions, stakeholder involvement may lead to studies that are misguided, cost too much, take too long, or are biased (Layzer's description of his 2004 Evaluators' Institute course; also see Cullen, Coryn, & Rugh, 2011).

Participatory and Collaborative Evaluations

- Participatory and collaborative approaches used for capacity building enhance program effectiveness and increase evaluation use (Patton's description of his 2004

Evaluators' Institute course; also see Cousins & Earl, 1992; Cousins & Whitmore, 1998).

- Political issues are especially present and influential in participatory evaluations (Brandon, 1998; House, 1993).

Use of Program Theory and Logic Models in Program Evaluations

- Positive effects of the use of program theory and logic models can include conceptual clarity of complex programs, motivation of staff, and better-focused evaluations (Funnel & Rogers, 2011; Rogers, 2008).
- Negative effects of the use of program theory and logic models can include diversion of time and attention from other critical evaluation activities, provision of an invalid or misleading picture of a program, and discouragement of critical investigation of causal pathways and unintended outcomes (Morell, 2010).
- Application of tried-and-true methods of using program theory and logic models helps evaluators and clients identify criteria, develop questions, and identify data sources and bases for comparisons (Donaldson, 2007).
- Inappropriate uses of program theory or logic models in evaluations include focusing only on intended outcomes, ignoring differential effects for individuals and client subgroups, and seeking only evidence that confirms the theory or model (Coryn, Noakes, et al. 2011).
- Effective strategies for avoiding use traps include the application of differentiated theory, market segmentation, and competitive elaboration of alternative hypotheses (Patricia Rogers's description of her 2004 Evaluators' Institute course; also see Birckmayer & Weiss, 2000).

Needs Assessment

- Appropriate uses of relevant needs assessments improve the relevance of conclusions about programs (James Altschuld's description of his 2004 Evaluators' Institute course; also see Stufflebeam, McCormick, Brinkerhoff, & Nelson, 1985).
- Reaching defensible judgments of the relevance and importance of identified program outcomes requires valid assessments of outcomes, treatments, and met and unmet needs (Stufflebeam, McCormick, et al., 1985; see also Altschuld & Witkin, 2000).
- Criteria and standards for use in program evaluations are more credible when grounded in needs assessments (Davidson, 2005; Scriven, 1991, 2007).

Evaluation of Program Implementation

- Effective evaluation of program implementation that yields feedback on critical ingredients of a program helps drive program improvement by fostering understanding of factors affecting variability in implementation and short-term results (Arnold Love's description of his 2004 Evaluators' Institute course; also see Cordray & Pion, 2006; Weiss, 1998).

Surveys

- Particular ways of developing, presenting, and encouraging responses to mail and Web survey questions contribute to high response rates and high-quality responses (Dillman, 2000; Dillman, Smyth, & Christian, 2009).
- Multiple sources of error must be overcome to produce high-quality survey results (Dillman, 2000; Dillman, Smyth, & Christian, 2009; also see Lord & Novick, 1968).
- Adhering to certain principles for writing survey questions minimizes measurement error (Dillman, 2000; Dillman, Smyth, & Christian, 2009; also see American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).
- Survey questions ordered in different ways have different, sometimes predictable consequences (Dillman, 2000; Dillman, Smyth, & Christian, 2009; also see Sijtsma & Junker, 1996).
- Self-administered questionnaires and telephone interviews yield different results for particular reasons (Dillman, 2000; Dillman, Smyth, & Christian, 2009).
- Different survey page layouts influence people to read and answer questions differently for particular reasons (Dillman, 2000; Dillman, Smyth, & Christian, 2009; also see Fowler, 1995).

Sampling

- Careful use of sampling methods can save resources and often increase the precision and accuracy of evaluation findings (Henry, 1990).

Applied Measurement

- Proper measurements of a program's feasibility, relevance, and effectiveness—that are systematic, replicable, interpretable, reliable, and valid—are necessary for successful evaluations (Ann Doucette's description of her 2004 Evaluators' Institute course; also see McDavid & Hawthorn, 2006).

Reporting Strategies

- Effective employment of a variety of reporting strategies beyond the written report, applied differentially to audiences, increases stakeholders' use of findings (Hallie Preskill's description of her 2004 Evaluators' Institute course; also see Coryn, 2006; Torres, Preskill, & Piontek, 2005).
- Appropriate use of design principles in reporting and presenting evaluation results increases usability (Evergreen, 2010).
- Identifying when, how, and to whom evaluation findings are to be disseminated and with what purpose enhances the effectiveness of evaluation reporting (Patton, 1997, 2008; Preskill & Torres, 1999a, 1999b).

Use of Technology in Evaluation

- Proper use of technology for data collection and analysis, storage and retrieval of information, and dissemination and use of findings contributes to strengthening

evaluations and reducing their costs (Love's description of his 2004 Evaluators' Institute course; also see Owen, 2006).

Building Organizational Capacity in Evaluation

- Developing and appropriately employing an organization's evaluation capacity lead to more and better learning in organizations (Preskill & Russ-Eft, 2005).

These hypotheses illustrate the need for empirical research on evaluation practices. Such research should be directed toward producing research-based principles for conducting program evaluations, akin to those found in the field of survey research. Dillman's research (2000; Dillman, Smyth, & Christian, 2009) on and development of survey methods provide an exemplar for emulation in other sectors of program evaluation. The groupings of the hypotheses just given also illustrate the range and complexity of dimensions to be considered in formulating program evaluation theories.

Potential Utility of Grounded Theories

In testing hypotheses about evaluation practices, it is important to document and take into account the subject program evaluation's particular circumstances, including pertinent contextual variables. Unlike laboratory experiments in the physical sciences, program evaluations typically occur in dynamic, uncontrolled settings; their procedures usually unfold in response to evolving stakeholder needs; and they are constrained and affected by complex and changing contextual circumstances. Explaining the functioning of a program evaluation requires extensive, valid description of the evaluation process, of the nature of contextual influences, and of the evaluation's impacts.

Accordingly, the program evaluation field could benefit by employing the methodology of grounded theories as one theory development tool. In applying this approach, theorists would generate theories grounded in systematic, rigorous documentation and analysis of actual program evaluations and their particular circumstances. This line of reasoning is consistent with a point made by Broom (1964) in his introduction to Kaplan's book *The Conduct of Inquiry*. Broom stated, "The behavioral scientist . . . needs to read from the strengths of his own understanding, insights, expertness, and subject matter and not from the insecurity of a limited familiarity with a remote discipline" (p. xvii). Strauss and Corbin (1990) described and illustrated grounded theory procedures that we see as potentially useful for generating sound theories of program evaluation.

Few if any examples of rigorously produced grounded theories of program evaluation have been published. However, the approaches examined in Parts Two and Three of this book comport with the general notion of grounded theory. They are prescriptions based on their authors' reflections on and critical analyses of a wide range of evaluation experiences. Limitations of these prescriptive theories are that they do not meet requirements for systematic and rigorous testing of theory-based hypotheses and that they lack documentation of the underlying program evaluation experiences.

Potential Utility of Metaevaluations in Developing Theories of Program Evaluation

A source of valuable evidence for use in developing program evaluation theories, akin to that from grounded theory work, is found in metaevaluations of program evaluations. These are studies that systematically document and assess program evaluations (a review and analysis of such studies can be found in Stufflebeam [2001c]). Program evaluators and their clients need to greatly increase their employment of metaevaluation and should make the results of such studies available to evaluation researchers. Fortunately, the *American Journal of Evaluation* encourages submission of metaevaluations. Researchers should use metaevaluation reports systematically to examine the reasons why different evaluation approaches succeeded or failed. We believe theory development efforts can profit from the use of metaevaluation findings to look at the adequacy and influence of guiding conceptualizations and procedures; implementation of the procedures; propriety considerations; stakeholder involvement; and contextual influences, including political forces and psychological factors.

Recent examples of using metaevaluations to formulate theoretical propositions about evaluation can be found in some of the doctoral dissertations completed at Western Michigan University, including Mafukidze-Trent's metaevaluation (2009) of HIV/AIDS prevention intervention evaluations in sub-Saharan Africa, Wingate's investigation (2009) of uses of *The Program Evaluation Standards* (Joint Committee, 1994) for metaevaluation, Sasaki's metaevaluation (2008) of several hundred international aid evaluations, Risley's metaevaluation (2007) of legislative program evaluations conducted by state legislatures in the United States, and Coryn's metaevaluation (2007) of government systems used throughout the world for evaluating and funding scientific research.

Program Evaluation Standards and Theory Development

Chapter 3 of this book examines in depth guiding principles and standards for evaluations, including, importantly, the Joint Committee's program evaluation standards (1981, 1994, 2011), which we refer to frequently. All three editions of *The Program Evaluation Standards* were designed, after intensive literature review and professional activities, to provide principles and guidelines for evaluating educational programs, projects, and materials in many different settings. Although focused on educational evaluation, the program evaluation standards have applicability and relevance in a wide range of professional and other arenas. This is evident in the Joint Committee's main requirements that evaluations be useful, feasible, proper, accurate, and accountable. Adding to the widespread applicability of the Joint Committee's program evaluation standards is the fact that all fields require evaluation of their training and educational programs.

The Joint Committee's 1994 and 2011 editions of *The Program Evaluation Standards* each comprise thirty individual standards. Each standard may be considered as a separate construct, thus open to empirical research. In addition, each standard includes guidelines to consider and apply as appropriate, plus common errors to avoid. These guidelines and common errors

essentially are hypotheses about what actions to take or avoid to conduct sound, effective evaluations. Although there is little doubt that the thirty standards individually and collectively have added considerable credibility and direction to the evaluation field, they will remain theoretical constructs until rigorously researched.

Let us consider, as an example, the third utility standard, Information Scope and Selection, which states, "Information collected should be broadly selected to address pertinent questions about the program and be responsive to the needs and interests of clients and other specified stakeholders" (Joint Committee, 1994, p. 37).[1] This standard underlines the importance of gathering a broad scope of relevant information that will meet all clients' decision-making objectives while also being sufficiently comprehensive for use in assessing an evaluand's merit and worth. In particular, the standard requires evaluators to assess a program "in terms of all important variables" (Joint Committee, 1994, p. 38); possible variables are effectiveness, harmful side effects, costs, responses to participants' needs, and the relevance of underlying assumptions and values. These aspects, taken together with the standard's stakeholder-centered orientation, impose the elements of a theory—but one in need of validation.

There undoubtedly are rich possibilities for empirical research based on all thirty standards and their associated guidelines and common errors. Directly and indirectly, the effective use of the program evaluation approaches explicated in Part Three of this book depends on satisfying the requirements of *The Program Evaluation Standards* (Joint Committee, 1994, 2011). Clearly, all sound evaluation approaches should meet conditions of utility, feasibility, propriety, accuracy, and evaluation accountability. Moreover, in the opinion of many evaluators, these standards provide not only useful guidelines but also established principles. Only research into these theoretical constructs will confirm their validity as predictors of evaluation outcomes and their standing as validated hypotheses. The stronger the links are between program evaluation approaches and practices, on the one hand, and standards for program evaluation, on the other, the more essential it is that assumptions contained in the latter are confirmed by research.

Summary

Sound theories of evaluation are needed to advance effective evaluation practices. An evaluation theory is different from, and more demanding in its requirements for validation than, the evaluation models and approaches presented later in this book. Program evaluations are part of a broad set of other types of evaluations (for example, of personnel, products, policies, and organizations). The history of formal program evaluation includes theoretical approaches that have proved useful, limited, or in some cases counterproductive (for instance, objectives-based evaluation and randomized controlled experiments). The definition of an evaluation theory is more demanding than that of an evaluation model (an evaluation theorist's idealized conceptualization for conducting program evaluations). An evaluation theory is defined as a coherent set of conceptual, hypothetical, pragmatic, and ethical principles forming a general framework to guide the study and practice of program evaluation. Beyond meeting these requirements, an evaluation theory should meet the following criteria: utility in efficiently generating verifiable predictions or propositions concerning evaluative acts and consequences; provision

of reliable, valid, actionable direction for ethically conducting effective program evaluations; and contribution to an evaluation's clarity, comprehensiveness, parsimony, resilience, robustness, generalizability, and heuristic power. Despite these demanding requirements of sound evaluation theories, theory development must be respected as a creative process that defies prescriptions of how to develop a sound theory.

The program evaluation field saw great theoretical progress in the last four decades of the twentieth century. Although much more work is needed, the field's literature is rich in concepts, standards, guiding principles, practical guidelines, and approaches. It is modestly strong in positing hypotheses, being strong in testing hypotheses about uses of surveys, but otherwise weak in presenting tested hypotheses. Overall, the program evaluation field has far to go in the quest to develop and present research-based theories whose predictions hold true.

The program evaluation field would benefit if future program evaluation theory development efforts would convert the best of the current program evaluation approaches into validated theories. In Part Two of this book, we give our assessment of which program evaluation approaches most merit serious theoretical development and practical use. The methods of grounded theories and information from metaevaluations of program evaluations could aid the needed theory development efforts. Moreover, the Joint Committee's *Program Evaluation Standards* (1994, 2011) provides a framework and hypotheses—in the form of standards, procedural guidelines, and common errors to avoid—to guide empirical research on evaluation.

REVIEW QUESTIONS

1. Argue the pros and cons of investing time and resources to develop and validate a sound theory of program evaluation.
2. Is it important to distinguish program evaluation theories from theories of other areas of evaluation, such as personnel evaluation? Why or why not?
3. What are the pros and cons of considering such historical evaluation approaches as objectives-based evaluation and randomized controlled experimentation as sufficient theories to guide evaluation work?
4. How would you explain to a client the distinction between an evaluation theory and an evaluation model?
5. Draft a checklist of criteria for use in evaluating evaluation theories.
6. Explain and assess the claim that theory development is a creative, arbitrary process.
7. Explain and assess the claim that multiple theories of program evaluation can be equally defensible.
8. Provide examples of conceptual, hypothetical, pragmatic, and ethical principles of program evaluation.

9. Outline a study you would conduct to test the hypothesis that appropriate application of the Joint Committee program evaluation standards enhances an evaluation's impacts.
10. Provide examples of how using information from metaevaluations would aid the development of a sound program evaluation theory.

Group Exercises

Work through the following two exercises with your group. It is quite possible that members will reach different conclusions about best methods to solve the problems. Members should try, however, to justify their point of view.

Exercise 1

A long-standing difference of opinion exists between two college faculties, the education faculty and humanities faculty, over the place of theory in evaluation practice, particularly as it applies to program evaluation. The education faculty contends that a logical start for students studying evaluation is to develop a solid theoretical foundation by reading about alternative evaluation approaches. The humanities faculty opposes this view, believing that evaluation is a pragmatic activity and that although a grasp of theory may develop over time, learning and applying a selected evaluation approach constitute the best way to begin. What advice do you give to the staff of these two faculties?

Exercise 2

Let one group member outline a program evaluation with which he or she is conversant. Now refer to the "Hypotheses for Research on Program Evaluation" section in this chapter, and select three hypotheses pertaining to different aspects of that evaluation. For instance, your group might select a hypothesis that could explain why the stakeholders in your case respected and used the evaluation's findings or, on the negative side, why many stakeholders failed to return questionnaires.

Note

1. We chose to reference the Joint Committee's 1994 version of the standard on the selection of information for an evaluation because that rendition stresses the importance of both addressing stakeholders' need for relevant information and obtaining sufficient information to judge a program's value. We are disappointed in what we see, in the Joint Committee's 2011 edition, as the watering down of the information requirements standard: although it stresses serving users' information needs. In our judgment, the 2011 version of the standard underplays the importance of also gathering whatever information is essential to judge a program's merit and worth.

Suggested Supplemental Readings

Alkin, M. C. (Ed.). (2004). *Evaluation roots: Tracing theorists' views and influences*. Thousand Oaks, CA: Sage.

Henry, G. T., & Mark, M. M. (2003). Toward an agenda for research on evaluation. In C. A. Christie (Ed.), *The practice-theory relationship in evaluation* (pp. 69–80). New Directions for Evaluation, no. 97. San Francisco, CA: Jossey-Bass.

Mark, M. M. (2007). Building a better evidence base for evaluation theory: Beyond general calls to a framework of types of research on evaluation. In N. L. Smith & P. R. Brandon (Eds.), *Fundamental issues in evaluation* (pp. 111–134). New York, NY: Guilford Press.

Miller, R. L. (2010). Developing standards for empirical examinations of evaluation theory. *American Journal of Evaluation*, *31*, 390–399.

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*, 1–19.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Thousand Oaks, CA: Sage.

Smith, N. L. (1993). Improving evaluation theory through the empirical study of evaluation practice. *Evaluation Practice*, *14*, 237–242.

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

Stufflebeam, D. L. (2001). The metaevaluation imperative. *American Journal of Evaluation*, *22*, 183–209.

Stufflebeam, D. L., & Shinkfield, A. J. (2007). *Evaluation theory, models, and applications*. San Francisco, CA: Jossey-Bass.

CHAPTER 3

STANDARDS FOR PROGRAM EVALUATIONS

Professional standards for guiding and judging evaluations are essential to the sound practice of evaluation. Such standards help ensure that evaluators and their clients communicate effectively and reach a clear, mutual understanding concerning the criteria an evaluation should meet, and they provide authoritative guidance for meeting the criteria. Professional standards also help prevent the possibility that either stakeholders or evaluators might unscrupulously bend evaluation outcomes to suit their own interests. When standards that define acceptable evaluation service are not met, the credibility of evaluation procedures, outcomes, or reporting is left in doubt. To be authoritative and credible, evaluation standards (1) must reflect a general consensus by experts in the conduct and use of evaluation, who were appointed by an adequate range of professional organizations representing users of evaluation reports and the technical specialties in evaluation, and (2) ideally should be approved by a body that accredits professional standards and standard-setting bodies.

During the past three decades, evaluators have considerably strengthened the professionalization of their emerging field by following the example of more mature fields and developing and using standards to guide and assess their evaluations. During this time, professional standards, directed toward sound practice through agreed-on principles, have become an integral part of the wider community's insistence on criteria and measures to ensure the quality, utility, fairness, and accountability of evaluations.

In this chapter we summarize and suggest ways to use the 2011 revision of *The Program Evaluation Standards,* developed by the Joint Committee on Standards for

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **The rationale and need for professional standards for program evaluations**
- **The three main sets of standards for program evaluators and evaluations**
- **How to apply these standards in program evaluation practice**

Educational Evaluation and accredited by the American National Standards Institute (ANSI); *Guiding Principles for Evaluators,* developed and officially endorsed by the American Evaluation Association (AEA; 2004); and the 2007 revision of *Government Auditing Standards,* developed by the U.S. Government Accountability Office (GAO) and employed in auditing U.S. government programs.

The three standard-setting bodies have published their somewhat unique definitions of an evaluation or auditing standard. The Joint Committee (2011) defined an evaluation standard as a "principle commonly agreed to by experts in the conduct and use of evaluation, that when implemented will lead to greater evaluation quality" (p. 292). In discussing this definition, the Joint Committee emphasized two features. First, a standard identifies and defines quality and guides evaluators and users in the pursuit of quality. Second, each Joint Committee standard is a voluntary consensus statement reflecting stakeholder input and Joint Committee deliberation and finalization in accordance with ANSI requirements, but it is not a law. The Joint Committee noted further that rather than specifying exact procedures, the 2011 standards require responsiveness and judgment in each evaluation setting. AEA (2004) noted that an evaluation principle provides evaluators with guidance that is general and conceptual rather than operational. GAO (2007) defined auditing standards as broad statements of auditors' responsibilities. Fundamentally, all three of these organizations' documents present general principles, which is the essential meaning of a standard. To aid communication, throughout this chapter we will use the generic term *standards* to refer to those in *The Program Evaluation Standards* (Joint Committee, 2011); AEA's *Guiding Principles for Evaluators* (2004); and *Government Auditing Standards* (U.S. Government Accountability Office, 2007).

All three documents provide authoritative direction for guiding and assessing program evaluation studies. *The Program Evaluation Standards* (Joint Committee, 2011) concentrates on evaluations of educational and training programs and services in the United States and Canada. It contains considerable specificity about what to do and not do in educational program evaluations and a range of illustrative cases. Although this standards document does not directly address evaluations in a wide range of substantive areas—such as engineering, philanthropy, social work, public administration, and community development—it is noteworthy that every field relies heavily on the education of its members and that its educational programs should be evaluated regularly. The program evaluation standards thus have applicability to educational enterprises in all fields. Moreover, a wide range of fields outside education and across the globe have applied the Joint Committee standards to program evaluations. AEA's *Guiding Principles for Evaluators* (2004) cuts across evaluations in many disciplines and service areas, but it lacks the detail and examples found in *The Program Evaluation Standards*. The U.S. Government Accountability Office's *Government Auditing Standards* (2007) is based on standards of the U.S. accounting and auditing professions; is focused on assessing and strengthening the effectiveness and financial accountability of U.S. agencies and federal programs; and contains detailed requirements, recommendations, and prohibitions. It is noteworthy that the GAO standards provide direction for evaluating audit organizations as well as the audits they conduct. This makes the government auditing standards potentially useful for assessing both program evaluations and the organizations that conduct these evaluations. Although the government

auditing standards concentrate on financial accounting requirements, this chapter focuses on the parts of these standards that provide direction for evaluating the goals, organization, implementation, and outcomes of not only federal programs but also departments and organizations that conduct program evaluations.

The three sets of standards treated in this chapter are among the most fully developed groupings of professional imperatives for program evaluations. The standards merit serious consideration by professional evaluators and those training to become evaluators in North America, and could be generally instructive to such persons in other countries. The standards level the playing field for neophyte and experienced evaluators by providing objective criteria for judging evaluation efforts and evaluation organizations. When armed with a working knowledge of the standards, evaluators can stand on firm ground in crafting and defending their program evaluation plans and assessments of evaluations, helping constituents understand what is required in sound program evaluations, and examining and strengthening an evaluation organization. Use of the standards documents as planning guides makes the design of evaluations a more certain and efficient process. When used in evaluations of program evaluations (that is, metaevaluations), they are especially important in helping users determine how much confidence to place in a given evaluation. Although the discussion in this chapter is particularly applicable to program evaluations in North America, program evaluators around the world may find these constructions instructive in regard to the pervasive issues in program evaluations that cross national boundaries as well as alternative formats and processes they might consider when crafting their own standards.

In this chapter we look at four topics. First, we discuss the need for evaluation standards and their functions. After that, we provide background on why and how the various North American evaluation standards were developed. Subsequently, we summarize each set of evaluation standards. Finally, we suggest ways of applying the different sets of standards. We hope readers will become allies in disseminating information about and productively using all three sets of evaluation standards discussed here. A great deal of work needs to be done to help evaluators and their clients learn about and effectively apply the evaluation standards. It is noteworthy that GAO maintains a Web site (www.gao.gov/govaud/ybk01.htm) of up-to-date information about the government auditing standards (U.S. Government Accountability Office, 2007) and provides many workshops on the standards. Also, AEA's Web site (www.eval.org/gptraining/gptrainingoverview.asp) offers a package for training users of the guiding principles for evaluators (AEA, 2004). We need to stress that mastering this chapter's content is only one important step toward developing a working knowledge of the subject standards. We urge readers to obtain, study, and apply the actual standards.

The Need for Evaluation Standards

Most professions and many other public service fields have developed and periodically update standards, principles, or codes of performance. They do so in the interest of having their members provide competent, ethical, and safe services. Often the standards, principles, or codes are part of an accrediting, licensing, or certification system intended to ensure high-quality services

and protect the interests of the public. Such standards, principles, or codes typically are defined by a standing committee of distinguished members of the service area, in some cases by government licensing or oversight bodies, and occasionally with participation by constituent and client groups. Familiar examples are the standards of practice employed by the fields of law, medicine, dentistry, hospitals, clinical psychology, engineering, educational and psychological testing, auditing, and accounting. Other examples are the codes established for the construction, electrical, plumbing, and food service areas. An important matter in advancing the development and use of standards is that periodically they must be reviewed and revised to keep them up to date, legally viable, and responsive to needs in the field.

We believe that every professional evaluator should know, understand, and faithfully apply appropriate standards of professional evaluation practice. Standards and codes are established and applied in the interest of ensuring and improving quality and protecting the public from shoddy, harmful, fraudulent, or wasteful evaluation services. Standards for program evaluations have several specific functions:

- Providing general principles for addressing a variety of practical issues in evaluation work
- Helping ensure that evaluators will employ the evaluation field's best available practices
- Providing direction to make evaluation planning efficient and inclusive of pertinent evaluation questions
- Providing core content for training and educating evaluators and other participants in the evaluation process
- Presenting evaluators and their constituents with a common language to facilitate communication and collaboration
- Helping evaluators achieve and maintain credibility among other professionals
- Helping evaluators earn and maintain credibility with public oversight bodies and clients
- Helping evaluators earn and maintain the public's confidence in the evaluation field
- Protecting consumers and society from harmful or corrupt practices
- Providing objective criteria for assessing and strengthening evaluation services
- Providing a basis for accountability by evaluators
- Providing a basis for adjudicating claims of malpractice and other disputes
- Providing a conceptual framework and working definitions to help guide research and development in evaluation

Adherence to professional standards for evaluations is at the very heart of professionalism and delivery of sound, useful evaluation services. We believe that current statements of professional standards for evaluations can serve these functions in the evaluation field.

The standards presented in this chapter were systematically developed, possess strong credibility, and are periodically reviewed and updated. The three sets of standards are distinct but also complementary. Learning and developing the facility to apply the three different sets

of standards selectively will enhance one's professionalism and versatility in conducting sound evaluations. Evaluators who are armed with a repertoire of alternative sets of standards are aided in conducting standards-based evaluations in a wide range of disciplines and service areas. Sometimes it will be appropriate to choose one set of standards over the others, because the set is compatible with the particular program area and preferred or mandated by the client group or oversight body. Even then it is often advantageous to derive guidance from two or even all three sets of standards. We can stand behind this position because all three sets of standards, in general, are in accord with the same fundamental principles of sound evaluation. One qualification concerning this point is that—in the wake of serious improprieties in the financial auditing sector, such as the famous case involving Arthur Andersen and Enron, and the ensuing federal Sarbanes-Oxley Act of 2002 to ensure fiscal accountability in corporations—the GAO standards place stronger emphasis on ensuring the independence of evaluations than do the AEA and Joint Committee standards.[1] We invite readers to study the following material on standards and incorporate what they find as valuable into their working philosophies of program evaluation.

Background of Standards for Program Evaluations

Program evaluators historically had no need to be concerned about explicit professional standards for program evaluations, because until relatively recently there was no semblance of an evaluation profession, and there were no standards for evaluations. Such standards came into prominence only during the 1980s and 1990s. Federal agencies had funded thousands of program evaluations as part of President Lyndon Johnson's War on Poverty and generally found them to be costly and poor in quality and utility. Efforts to reform this embryonic program evaluation movement included establishing authoritative standards and principles for assessing and strengthening evaluation plans and reports. The development of the initial sets of standards and principles signaled both the evaluation field's immaturity and weakness and an added step in its movement toward professionalization.

With the evolution of evaluation as a profession becoming a reality only during the last quarter of the twentieth century, there was a growing sense among practitioners that acceptable codes of evaluator behavior were needed. The Joint Committee was established in 1975. Through the years, this standing committee has continued to be sponsored by twelve to seventeen professional societies with a combined membership totaling nearly three million. The committee's charge is to perform ongoing development, reviews, and revisions of standards for educational evaluations. This committee issued *Standards for Evaluations of Educational Programs, Projects, and Materials* in 1981; an updated version in 1994, *The Program Evaluation Standards;* and the third edition in 2011. The Joint Committee also published standards for evaluating education personnel in 1988 and 2009, and in 2003 it issued a set of standards for evaluations of students. The Joint Committee is accredited by ANSI as the only body recognized to set standards for educational evaluations in the United States. Its members are from Canada as well as the United States, and its current standards are intended for use throughout North America.

At nearly the same time as the first Joint Committee standards were published, the Evaluation Research Society (ERS) produced a second set (Evaluation Research Society Standards Committee, 1982). ERS, established in 1976, focused on professionalizing program evaluation as practiced across a wide range of disciplines, government programs, and service areas. ERS's standards for program evaluation were fifty-five brief, admonitory statements divided into the following categories: (1) formulation and negotiation, (2) structure and design, (3) data collection and preparation, (4) data analysis and interpretation, (5) communication and disclosure, and (6) use of results. In 1986 ERS amalgamated with the Evaluation Network (ENet) to form AEA, which currently has a membership of nearly six thousand (see Chapter 1). AEA subsequently retired the ERS standards in favor of producing its own guiding principles (Shadish, Newman, Scheirer, & Wye, 1995b). In July 2004 AEA members ratified a revised edition of *Guiding Principles for Evaluators*.

The U.S. General Accounting Office (later renamed the U.S. Government Accountability Office) explicitly included program auditing in its 2002, 2003, and 2007 editions of *Government Auditing Standards*. President Johnson's War on Poverty, which began in 1965, spawned many expensive federal programs, which generated a huge need for financial auditing of these programs. In 1972 the U.S. General Accounting Office began issuing government auditing standards. The initial edition and early revisions of the document containing these standards dealt almost exclusively with the financial aspects of federal programs. The 2003 and 2007 editions include program audits as one of the foci of general standards and present chapters containing fieldwork standards and reporting standards for program performance audits.

Especially noteworthy in the 2003 and 2007 editions is the section on independence, which prohibits auditors from simultaneously providing both auditing and consulting services to the same entity. Such commingling of services is seen as an unacceptable conflict of interest. It could lead auditors to evaluate their own work and thus lose their independence and credibility, possibly succumbing to illicit pressures to distort reports. This pitfall has been made manifestly clear in the private sector. For example, Arthur Andersen auditors both consulted with and audited the work of Enron, and Andersen was later charged with covering up and being party to Enron's malfeasance. Andersen's alleged compromise of its independence contributed to the scandal in which Enron's employees and stockholders lost billions of dollars. Ultimately, this transgression led to the near demise of Andersen, previously one of America's Big Five auditing firms. Clearly, the GAO (2007) standard on independence is applicable to program evaluations as well as financial audits, and we think its message probably should be incorporated more strongly into future editions of both the Joint Committee's *Program Evaluation Standards* and AEA's *Guiding Principles for Evaluators*.

Joint Committee Program Evaluation Standards

The seventeen members of the original Joint Committee were appointed by twelve professional organizations. The organizations and their appointed members represented a wide range of specialties: school accreditation, counseling and guidance, curriculum development,

educational administration, higher education, educational measurement, educational research, educational governance, program evaluation, psychology, statistics, and teaching. A fundamental requirement of the committee is that it include about equal numbers of members representing client and evaluator perspectives.[2] Over the years, the number of organizations sponsoring the Joint Committee has increased. (At the publication of the 2011 edition of *The Program Evaluation Standards,* the committee was sponsored by seventeen organizations, including AEA.[3]) The Joint Committee's work was housed at the Western Michigan University Evaluation Center from 1975 to 2009, and since then has been located at the University of Iowa.

Each edition of *The Program Evaluation Standards* (Joint Committee, 1981, 1994, 2011) has detailed presentations of thirty standards. Each standard contains a statement of the standard, a rationale for and explanation of its requirements, guidelines for carrying it out, common errors or hazards to be anticipated and avoided, an illustrative case, and supporting documentation. The 1994 and 2011 versions cover education and training in such settings as business, government, law, medicine, the military, nursing, professional development, elementary and secondary schools, social service agencies, and colleges and universities.

Whereas the thirty Joint Committee standards in the 1981 and 1994 editions are grouped according to four essential attributes of a sound evaluation—utility, feasibility, propriety, and accuracy—the 2011 edition includes a fifth attribute—evaluation accountability. The 2011 edition of *The Program Evaluation Standards* advises both evaluators and clients to apply the thirty standards so that their evaluations satisfy all five essential attributes of a sound evaluation. We advise readers to fix firmly in their minds the following five fundamental concepts presented in the 2011 version of *The Program Evaluation Standards*: utility, feasibility, propriety, accuracy, and evaluation accountability.

Utility

The utility standards are intended to ensure that an evaluation effectively delivers information and judgments that stakeholders can apply to such areas as program planning, control, improvement, assessment, accountability, and dissemination. An evaluation should be useful. It should be addressed to those persons and groups that are involved in or responsible for implementing the program being evaluated. The evaluator should ascertain the users' information needs and report to them relevant evaluative feedback that is clear, concise, and on time. He or she should help them identify and attend to the program's problems and be aware of important strengths. The evaluator should address the users' most important questions while also obtaining the full range of information needed to assess the program's merit and worth. Finally, the evaluator should not only report feedback about strengths and weaknesses but also help users study and apply the findings. The utility standards reflect the general consensus found in the evaluation literature that program evaluations should effectively address the information needs of clients and other right-to-know audiences and should inform program improvement processes and program accountability reports. If there is no prospect that the findings of a contemplated evaluation will be used, the evaluation should not be undertaken.

Feasibility

An evaluation should be feasible. The evaluator should employ evaluation procedures that are parsimonious and operable in the program's environment, should avoid disrupting or otherwise impairing the program, and should control as much as possible the political forces that might otherwise impede or corrupt the evaluation. And the evaluation should be conducted as efficiently and cost-effectively as possible. This set of standards emphasizes that evaluation procedures must be workable in real-world settings, not only in experimental laboratories. Overall, the feasibility standards require evaluations to be realistic, prudent, diplomatic, politically viable, frugal, and cost effective. Despite federal mandates to the contrary, true experiments should not be applied in field settings, where it is impossible to meet this approach's required assumptions. Instead, it is often prudent to conduct a naturalistic, multimethod study, such as an in-depth case study.

Propriety

An evaluation should meet conditions of propriety. It should be grounded in clear, written agreements defining the obligations of the evaluator and client in regard to supporting and executing the evaluation. The evaluator should protect all involved parties' rights and dignity, and the evaluation's findings must be honest and not distorted in any way. Reports should be released in accordance with advance disclosure agreements and applicable freedom of information statutes. Moreover, reports should convey appropriately balanced accounts of strengths and weaknesses. The propriety standards reflect the fact that evaluations can affect many people in negative as well as positive ways. They are designed to protect the rights of all parties to an evaluation. In general, the propriety standards require that evaluations be conducted legally, ethically, and with due regard for the welfare of those involved in the evaluation as well as those affected by the results.

Accuracy

The accuracy standards are intended to ensure that an evaluation will yield, and the evaluator will convey, technically adequate information about the features that determine the merit and/or worth of the program being evaluated. The evaluator should clearly describe the program as it was planned and actually executed, describe the program's background and setting, and report valid and reliable findings. He or she should identify and substantiate the appropriateness of the evaluation's information sources, measurement methods and devices, analytical procedures, and provisions for bias control. The evaluator should present the strengths, weaknesses, and limitations of the evaluation's plan, procedures, information, and conclusions, and should describe and assess the extent to which the evaluation provides an independent, unbiased assessment as opposed to a possibly biased self-assessment. In general, this group of standards requires evaluators to obtain technically sound information, analyze it correctly, report justifiable conclusions, and note any pertinent caveats. The overall rating of an evaluation against the accuracy standards is an index of its overall validity.

Evaluation Accountability

An evaluation should be fully accountable. The evaluator should document and make available for inspection all aspects of the evaluation that are needed for independent assessments of the evaluation's utility, feasibility, propriety, accuracy, and accountability. The evaluator should also conduct an internal assessment of the evaluation and attest to the extent to which it meets all of the standards. In addition, the evaluator should be proactive in seeking an independent, standards-based assessment of the evaluation—that is, an external metaevaluation; he or she should cooperate throughout the process and advocate release of metaevaluation findings.

The Joint Committee's Overall Approach

These five concepts are the foundation stones in the 2011 Joint Committee standards. In addition to the 1981, 1994, and 2011 editions of *The Program Evaluation Standards,* the committee developed *The Personnel Evaluation Standards* (1988, 2009) and *The Student Evaluation Standards* (2003). In each of its standard-setting projects, the Joint Committee engaged about two hundred people concerned with the professional practice of evaluation in a systematic process of generating, testing, and clarifying widely shared principles by which to guide, assess, and govern evaluation work in education. In each project, the committee sought widely divergent views on what standards should be adopted and subjected draft standards to field tests and national hearings. The committee subsequently worked through consensus development processes to converge on the final set of standards.

Each set of standards released by the Joint Committee is a living document. This standing committee encourages users of each set of standards to provide feedback on applications of the standards, along with criticisms and suggestions. From the outset of its work, the Joint Committee has provided for periodic reviews and improvement of the standards. This feature of how it operates is consistent with requirements for maintaining its accreditation by ANSI.

The Joint Committee's 2011 program evaluation standards are summarized in Exhibit 3.1. ANSI approved these standards as an American National Standard on June 21, 2010. Readers are advised to study the full text of *The Program Evaluation Standards,* in the interest of internalizing the standards and applying them judiciously at each stage of an evaluation. The summary presented in Exhibit 3.1 is only a starting point and convenient memory aid.

Exhibit 3.1 SUMMARY OF *THE PROGRAM EVALUATION STANDARDS*

Utility Standards

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 Identification. Evaluations should clarify and specify the individual and cultural values underpinning purposes, processes, and judgments.

U5 Relevant Information. Evaluation information should serve the identified and emergent needs of stakeholders.

U6 Meaningful Processes and Products. Evaluations should construct activities, descriptions, and judgments in ways that encourage participants to rediscover, reinterpret, or revise their understandings and behaviors.

U7 Timely and Appropriate Communicating and Reporting. Evaluations should attend to the continuing information needs of their multiple audiences.

U8 Concern for Consequences and Influence. Evaluations should promote responsible and adaptive use while guarding against unintended negative consequences and misuse.

Feasibility Standards

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 Standards

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 Standards

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 Standards

E1 Evaluation Documentation. Evaluations should fully document their negotiated purposes and implemented designs, procedures, data, and outcomes.

E2 Internal Metaevaluation. Evaluators should use these and other applicable standards to examine the accountability of the evaluation design, procedures employed, information collected, and outcomes.

E3 External Metaevaluation. Program evaluation sponsors, clients, evaluators, and other stakeholders should encourage the conduct of external metaevaluations using these and other applicable standards.

Source: Joint Committee on Standards for Educational Evaluation. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

The Joint Committee (2011) offered advice on which of the thirty standards are most applicable to each of ten tasks in the evaluation process: deciding whether to evaluate; negotiating and formalizing agreements, contracts, and budgets; determining who will evaluate; negotiating and developing evaluation purposes and questions; describing the program; designing the evaluation; managing the evaluation; collecting information, analyzing information; and communicating and reporting findings. The committee's judgments of the different standards' applicability to each evaluation task are summarized in Table 3.1. The categories of standards are listed across the top of the matrix, and the ten evaluation tasks are presented down the side. Although the Joint Committee concluded that all of the standards are applicable in all educational program evaluations, the various cells indicate the committee's judgment of which standards are likely to be most relevant to given evaluation tasks. The 2011 program evaluation standards are particularly applicable in metaevaluations. In such studies, the metaevaluator collects information and judgments about the extent to which a program evaluation complied with the requirements for meeting each standard. The metaevaluator then judges whether each standard was addressed, partially addressed, not addressed, or not applicable. A profile of these judgments provides a basis for judging the evaluation against the considerations of utility, feasibility, propriety, accuracy, and evaluation accountability and in relation to each standard. When such metaevaluations are carried out early in an evaluation, they provide diagnostic feedback of use in strengthening the evaluation. When completed after a program evaluation, a metaevaluation helps users assess the evaluation's findings and recommendations to determine whether to make prudent use of them or to reject them in part or even completely. (Checklists for applying the program evaluation standards are available from the Evaluation Center's Web site at www.wmich.edu/evalctr/checklists as well as this book's Web site at www.josseybass.com/go/evalmodels.)

American Evaluation Association Guiding Principles for Evaluators

In November 1992 AEA created a task force and charged it with developing general guiding principles for evaluation practice. The task force, chaired by William R. Shadish, subsequently drafted a set of guiding principles for evaluators (Shadish et al. 1995b). Following a review process made available to the entire AEA membership, the task force finalized the principles document. After an affirmative vote by the AEA membership, the AEA board adopted the task force's recommended principles as the official AEA evaluation principles. AEA then published the principles in *New Directions for Program Evaluation* in 1995 (Shadish et al., 1995b). From 2002 to 2003 the guiding principles were reviewed and revised by the AEA Ethics Committee. In July 2004 the AEA membership ratified the revised *Guiding Principles for Evaluators*. The 2004 AEA guiding principles comprise five principles and twenty-five underlying normative statements to guide evaluation practice, and can be accessed from the AEA Web site at www.archive.eval.org/Publications/GuidingPrinciples.asp. AEA gives blanket permission to reprint *Guiding Principles for Evaluators* with appropriate attribution. Exhibit 3.2 shows the principles and the associated normative statements as they appear regularly in the inside cover of issues of the *American Journal of Evaluation*.

Table 3.1 Analysis of the Relative Importance of the Five Categories of Program Evaluation Standards in Performing the Tasks in an Evaluation

	Utility Standards	Feasibility Standards	Propriety Standards	Accuracy Standards	Evaluation Accountability Standards
Deciding whether to evaluate	U1, U2, U3	F2, F3	P2, P3, P4, P6	A4	E1
Negotiating and formalizing agreements, contracts, and budgets	U1, U2, U6, U7	F1, F4	P2, P4, P6, P7		E2, E3
Determining who will evaluate	U1, U2, U4	F1, F3, F4	P2, P6		E1
Negotiating and developing evaluation purposes and questions	U2, U3, U4, U5, U6	F2, F3	P4, P5, P6	A1, A2, A4, A7	E1
Describing the program	U2	F2, F3	P1, P5, P6	A2, A3, A4, A7, A8	E1
Designing the evaluation	U2, U3, U4, U6	F2, F3, F4	P1, P2, P3	A1, A2, A3, A4, A5, A6, A7	E1
Managing the evaluation	U1, U2, U6, U7	F1, F2, F3	P7	A1, A6	E1
Collecting information	U2, U5, U6, U7, U8	F1, F2, F3	P3	A2, A3, A5	E1, E2
Analyzing information	U4, U5, U6, U7	F1, F2, F3	P1, P4, P5	A1, A2, A3, A4, A5, A6, A7, A8	
Communicating and reporting	U2, U5, U7, U8	F2, F3	P1, P3, P5, P6	A7, A8	

Source: Adapted from Joint Committee on Standards for Educational Evaluation. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Exhibit 3.2 AMERICAN EVALUATION ASSOCIATION GUIDING PRINCIPLES FOR EVALUATORS

A. *Systematic Inquiry*. Evaluators conduct systematic, data-based inquiries, and thus should:

   1. Adhere to the highest technical standards appropriate to the methods they use.
   2. Explore with the client the shortcomings and strengths of evaluation questions and approaches.
   3. Communicate the approaches, methods, and limitations of the evaluation accurately and in sufficient detail to allow others to understand, interpret, and critique their work.

B. *Competence*. Evaluators provide competent performance to stakeholders, and thus should:

   1. Ensure that the evaluation team collectively possesses the education, abilities, skills, and experience appropriate to the evaluation.

2. Ensure that the evaluation team collectively demonstrates cultural competence and uses appropriate evaluation strategies and skills to work with culturally different groups.
3. Practice within the limits of their competence, decline to conduct evaluations that fall substantially outside those limits, and make clear any limitations on the evaluation that might result if declining is not feasible.
4. Seek to maintain and improve their competencies in order to provide the highest level of performance in their evaluations.

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, and thus should:

1. Negotiate honestly with clients and relevant stakeholders concerning the costs, tasks, limitations of methodology, scope of results, and uses of data.
2. Disclose any roles or relationships that might pose a real or apparent conflict of interest prior to accepting an assignment.
3. Record and report all changes to the original negotiated project plans, and the reasons for them, including any possible impacts that could result.
4. Be explicit about their own, their clients', and other stakeholders' interests and values related to the evaluation.
5. Represent accurately their procedures, data, and findings, and attempt to prevent or correct misuse of their work by others.
6. Work to resolve any concerns related to procedures or activities likely to produce misleading evaluative information, decline to conduct the evaluation if concerns cannot be resolved, and consult colleagues or relevant stakeholders about other ways to proceed if declining is not feasible.
7. 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, and thus should:

1. Seek a comprehensive understanding of the contextual elements of the evaluation.
2. Abide by current professional ethics, standards, and regulations regarding confidentiality, informed consent, and potential risks or harms to participants.
3. Seek to maximize the benefits and reduce any unnecessary harms that might occur from an evaluation and carefully judge when the benefits from the evaluation or procedure should be foregone because of potential risks.

4. Conduct the evaluation and communicate its results in a way that respects stakeholders' dignity and self-worth.
5. Foster social equity in evaluation, when feasible, so that those who give to the evaluation may benefit in return.
6. Understand, respect, and take into account differences among stakeholders such as culture, religion, disability, age, sexual orientation, and ethnicity.

E. *Responsibilities for General and Public Welfare*. Evaluators articulate and take into account the diversity of general and public interests and values, and thus should:

1. Include relevant perspectives and interests of the full range of stakeholders.
2. Consider not only immediate operations and outcomes of the evaluation, but also the broad assumptions, implications, and potential side effects.
3. Allow stakeholders' access to, and actively disseminate, evaluative information, and present evaluation results in understandable forms that respect people and honor promises of confidentiality.
4. Maintain a balance between client and other stakeholder needs and interests.
5. Take into account the public interest and good, going beyond analysis of particular stakeholder interests to consider the welfare of society as a whole.

Source: American Evaluation Association. (2004). *Guiding principles for evaluators*. Washington, DC: Author. Retrieved from http://www.archive.eval.org/Publications/GuidingPrinciples.asp

The 2004 AEA guiding principles for evaluators provide evaluators with a code of professional behavior. The principles are also applicable to evaluating evaluation designs and reports across a wide array of disciplines. They encourage evaluators to practice systematic inquiry and to serve society by acting honestly and giving priority to the public welfare throughout their professional career and in conducting evaluations.

Government Auditing Standards

David M. Walker, comptroller general of the United States, released the 2007 revision of *Government Auditing Standards*, a document commonly referred to as Generally Accepted Government Auditing Standards (GAGAS), on behalf of the U.S. Government Accountability Office. A GAO project team headed by Jeffrey C. Steinhoff had developed the document through a deliberative process including extensive public comments and input from the comptroller general's twenty-one-member Advisory Council on Government Auditing Standards, chaired by Jack R. Miller. This revision incorporates the then-current fieldwork and reporting standards issued by the American Institute of Certified Public Accountants. GAGAS applies to

the work of both individual auditors and audit organizations. Later in this section we provide verbatim statements of standards in GAGAS and then paraphrase GAO's narrative concerning the standards.

GAGAS advances auditing of government programs as vital to fulfilling the government's duty to be accountable to legislative bodies, government officials, and the people. Auditors are seen as responsible for helping interested parties assess and ensure the validity of reported information on the results of programs and the soundness of related systems of internal control. GAGAS presents broad statements of auditors' responsibilities that are intended to represent a floor of acceptable auditing behavior. They require auditors to meet defined requirements for independence, professional judgment, competence, audit quality control and assurance, and external peer reviews in planning, conducting, and reporting on their work. Auditors and audit organizations are expected to follow these standards when required by law, regulation, contract, agreement, or policy. The standards provide a framework to ensure that audits are valid and relevant and also instrumental in improving government management, decision making, oversight, effectiveness, efficiency, and accountability. These standards require auditors to serve their clients and other financial statement users and to protect the public interest by scrutinizing internal controls and reporting on the extent to which the controls deter fraudulent financial reporting, illegal acts, or abuses; protect assets; and provide an early warning of emerging problems. Auditors who perform work in accordance with GAGAS are expected to justify any departures from these standards.

The document has eight chapters: (1) "Use and Application of the *Government Auditing Standards*"; (2) "Ethical Principles in Government Auditing"; (3) "General Standards"; (4) "Field Work Standards for Financial Audits"; (5) "Reporting Standards for Financial Audits"; (6) "General, Field Work, and Reporting Standards for Attestation Engagements"; (7) "Field Work Standards for Performance Audits"; and (8) "Reporting Standards for Performance Audits." Chapters 1, 3, 7, and 8 of GAGAS are particularly applicable to program evaluations.

Several points from Chapter 1 of GAGAS are especially pertinent to program evaluations. Targeted types of program audits vary widely and may include assessments of program effectiveness, side effects, economy, efficiency, cost-effectiveness, program implementation, and equitable distribution of resources and services; the reliability, validity, or relevance of an organization's performance measures; internal control; compliance with regulations; and prospective analyses to assist with program planning. Outcomes to be sought from program audits include program improvement, cost reduction, facilitation of decision making, and public accountability. Chapter 1 notes that GAGAS may be used in conjunction with other standards, including the AEA guiding principles for evaluators (2004) and the Joint Committee's program evaluation standards (1994).

Chapter 2 of GAGAS identifies and defines the key ethical principles underlying the standards, as summarized here:

The public interest: Honoring and protecting the collective well-being of the community of people and entities being served by the auditors

Integrity: Conducting work with an attitude that is objective, fact oriented, nonpartisan, and nonideological with regard to audited entities and users of the auditors' reports

Objectivity: Being independent in fact and appearance when providing audit services, maintaining an attitude of impartiality, being intellectually honest, and being free of conflicts of interest

Proper use of government information, resources, and position: Using government information, resources, or positions for official purposes and not inappropriately for the auditors' personal gain or in a manner contrary to law or detrimental to the legitimate interests of the audited entity or the audit organization; and sensitive or classified information is to be handled properly

Professional behavior: Complying with laws and regulations and avoiding any conduct that might bring discredit to auditors' work, including actions that could cause an objective third party with knowledge of the relevant information to conclude that the auditors' work was professionally deficient

In accordance with these principles, program audits or evaluations should be objective, competent, fact based, intellectually honest, nonpartisan, free of conflicts of interest (in fact or appearance), and nonideological in relationships with evaluees and users of audit reports. Auditors should honor the public trust, be professional in planning and performing their assessment and reporting functions, and embody the concept of accountability to the public. They should not use obtained information for any personal gain or in any manner that would impede the legitimate and ethical efforts of the audited entity.

GAGAS General Standards

As presented in Chapter 3, GAGAS contains four general standards having to do with (1) the independence of the audit organization and its individual auditors; (2) the exercise of sound professional judgment in conducting and reporting audits, exercising quality control, and engaging external peer reviews; (3) the competence and continuing education of audit staff; and (4) provisions for quality control to provide reasonable assurance of compliance with applicable auditing standards. These general standards are considered mandatory when performing audits requiring application of GAGAS. They are targeted to ensure credibility of audit results. Discussion of the general auditing standards follows.

Independence

The general standard related to independence reads as follows:

> In all matters relating to the audit work, the audit organization and the individual auditor, whether government or public, must be free from personal, external, and organization impairments to independence, and must avoid the appearance of such impairments of independence. (U.S. Government Accountability Office, 2007, p. 29)

This standard is intended to ensure that opinions, conclusions, judgments, and recommendations will be impartial and viewed as impartial by knowledgeable third parties, and be free from personal, external, and organizational impairments to independence.

Personal Impairments Audit organizations are expected to maintain internal quality control systems for detecting and appropriately addressing issues, in fact or appearance, of auditor partiality. Individual auditors are expected to notify appropriate officials within their audit organization if they have any personal impairments to independence—for example, friends or family members in the audited entity, financial interest in the entity, previous employment in the entity, seeking employment in the entity, or biases toward any aspect of the entity. To forestall or address personal impairments to independence, audit organizations are tasked with establishing pertinent independence policies and procedures, communicating these to all auditors in the organization, providing them with appropriate training, monitoring compliance with the standard, establishing a disciplinary mechanism, promptly resolving personal infringements of independence, and stressing that auditors must maintain independence and always act in the public interest.

External Impairments An auditor's independence may be compromised when factors outside the audit organization constrain or interfere with the auditor's ability to render independent and objective opinions and conclusions. Such impairments may occur when managers in the audited organization, oversight body, or funding organization exert pressure to inappropriately alter the audit's scope or procedures; unreasonably restrict time allowed to complete the audit; restrict access to needed records; interfere with the selection and appointment of auditors; unduly restrict funding of the audit; pressure the auditor to distort, excise, or tone down certain judgments; inappropriately jeopardize the auditor's continued employment; or threaten or actually engage in inappropriate modification of the audit report. Audit organizations are expected to identify possible external impairments and ways of addressing them through internal policies and procedures for reporting and resolving external impairments.

Organizational Impairments Audit organizations need to be free from impairments to independence regarding their place within or relationship to the organization that houses the entity to be audited. Auditors can be presumed to be free of organizational impairments if their audit organization is independent from the audited entity.

The Independence standard spells out a number of ways that audit organizations can meet the requirement of organizational independence. These include assignment of audit responsibility to a level of government other than the one housing the audited entity and assignment to a different branch of government within the same level of government as the audited entity. Also, an audit organization may be presumed to be free of organizational impairments to independence if its head was directly elected or appointed by a government entity that oversees and has the power to remove the person.

This standard identifies a number of safeguards that may be appropriate for audit organizations that are in structures different from those just referenced. Among these safeguards against organizational impairments to independence are statutory protections that prevent abolishment of the audit organization by the audited entity; require transparency of reasons for removing the head of the audit organization; prevent the audited entity from interfering

in the audit; require the audit organization to report to a governing body that is independent from the audited entity; give the audit organization sole authority over staffing the audit work; and guarantee access to records and documents needed to complete an audit.

Some audit organizations are internal to the organization being audited. To meet the needs of these internal audit organizations, the Independence standard includes provisions and suggestions that may apply to audit units within such organizations as colleges, universities, school districts, and hospitals. Main requirements for such units are accountability to the head or deputy head of the organization, reporting audit results to the head or deputy head of the organization and to those charged with governance, and being located organizationally outside the staff or line management function of the unit under audit. The internal audit organization must document the conditions that make it free from organizational impairments to independence and must subject this documentation to peer review to ensure that all necessary safeguards are met.

Professional Judgment

The second GAGAS general standard pertains to professional judgment as used in planning and performing audits and attestation engagements and in reporting results.[4] This standard requires auditors to apply reasonable care, a questioning mind, a critical assessment of evidence, and an honest appraisal of their own qualifications for performing the audit to all aspects of their work, to serve the public interest effectively and maintain utmost integrity, objectivity, independence, and credibility. The exercise of professional judgment is intended to help ensure that any material misstatements or significant inaccuracies in data will be detected, considered, minimized, mitigated, and explained.

Exercise of professional judgment is required in determining the type of assignment to be performed, pertinent evaluative criteria, scope of the work, methodological approach, type and amount of needed information, and tests and procedures. Professional judgment is also required in staffing the audit, carrying out the study, assessing obtained evidence, evaluating the audit work, and reporting findings.

Auditors are expected to maintain professional skepticism throughout an assignment in judging the sufficiency, competency, and relevance of evidence. They are not to assume that management at the audited entity is dishonest or unquestionably honest. Instead, they are expected to judge evidence on its merits, regardless of beliefs about the honesty and integrity of management.

Competence

The third GAGAS general standard requires that "the staff assigned to perform the audit or attestation engagement . . . collectively possess adequate professional competence for the tasks required" (U.S. Government Accountability Office, 2007, p. 51). This standard places responsibility on audit organizations to assign staff members who collectively possess the knowledge, skills, and experience needed to meet the needs of a particular assignment or set of audits to be performed. Audit organizations are expected to maintain a competent

workforce by employing a sound process of staff recruitment, hiring, continuous development, assignment, and evaluation. To meet the requirement for competence, an audit organization may have to employ persons with expertise in a variety of areas, such as statistical sampling, survey methods, statistical analysis tests, accounting, law, audit design and methodology, information technology, public administration, engineering, economics, social sciences, and actuarial science. Auditors are expected to maintain their professional competence through continuing professional education, in accordance with the GAGAS requirements.

In regard to work assignments, each staff member is expected to have relevant education, skills, and experience and be knowledgeable of the parts of GAGAS that pertain to their assignments. An audit team collectively should possess knowledge of the audited entity's environment and the subject matter under review, and should possess good oral and written communication skills.

Quality Control and Assurance

The fourth general standard, quality control and assurance, reads:

> Each audit organization performing audits or attestation engagements in accordance with GAGAS must: (a) establish a system of quality control that is designed to provide the audit organization with reasonable assurance that the organization and its personnel comply with professional standards and applicable legal and regulatory requirements, and (b) have an external peer review at least once every 3 years. (U.S. Government Accountability Office, 2007, p. 55)

This standard's intent is to make certain that an audit organization will have and implement a structure, policies, and procedures for internal quality control that will reasonably ensure the organization's compliance with GAGAS.

The internal quality control mechanism should include policies and procedures that (1) designate responsibility for ensuring the quality of audits and communicating policies and procedures relating to quality; (2) provide reasonable assurance that independence in audits will be maintained; (3) help see to it that the organization will initiate, accept, and continue only those audits that comply with professional standards, ethical principles, and pertinent legal authority; (4) help ensure that the organization will have appropriately qualified staff; (5) help ensure that audit assignments will be appropriately documented and reported; and (6) provide for ongoing, periodic assessment of the organization's compliance with quality control policies and procedures.

Internal quality control systems are permitted to vary in their sophistication and documentation depending on such factors as the size, nature, and resources of the audit organization and appropriate cost-benefit considerations. Nevertheless, each audit organization is expected to maintain up-to-date documentation that details and demonstrates compliance with its quality control policies and procedures.

As already indicated, audit organizations are expected to have an independent, external peer review of their auditing and attestation practices at least once every three years. Peer reviews

are to focus on the appropriateness, adequacy, and effective implementation of internal quality control policies and procedures. The peer review team's organization is required to meet all relevant conditions for independence as defined in GAGAS and ultimately this organization is expected to assign audit team members who collectively are knowledgeable of GAGAS and government auditing; are competent to conduct the particular review; and give evidence of having conducted competent peer reviews. Also, the audit organization should provide the entity to be audited with one or more peer review reports assessing the organization's auditing qualifications and experience.

The peer review team is expected to examine the audit organization's internal quality control policies and procedures and interview various staff members at the audit organization to assess their understanding of and compliance with relevant quality control policies and procedures. The team will also sample and assess a reasonable cross-section of the audit organization's performed assignments. The review should be sufficiently comprehensive to conclude whether the audit organization's quality control policies, procedures, and actions are sufficient to provide reasonable assurance that the organization's work conforms to appropriate standards. The peer review team's written report should include at least the scope of the review; caveats; the standards used to assess the audit organization's peer review system; an opinion on the adequacy of the structure and implementation of the organization's peer review system; and, as appropriate, reasons for an adverse opinion plus actionable recommendations.

The importance of peer reviews is seen in the requirement that audit organizations seeking to contract for work in accordance with GAGAS provide the party contracting for their services with their most recent external peer review report and associated materials. Audit organizations are also expected to transmit their external peer review report to appropriate oversight bodies. The organizations are advised, on request by government oversight bodies or other right-to-know groups, to promptly make public their peer review report and associated materials.

The four general standards of GAGAS are heavily oriented to audit organizations but also have considerable relevance to program evaluations and the organizations that conduct program evaluations. Clearly, program evaluations can be enhanced by meeting general requirements for independence, professional judgment, competence, and quality control and assurance. We next consider the two chapters on fieldwork and reporting in GAGAS that directly address the enterprise of program evaluation.

GAGAS Fieldwork Standards for Performance Audits

As portrayed in Table 3.2, the fieldwork standards for performance audits have two dimensions: (1) three general, pervasive concepts along the horizontal dimension and (2) four standards along the vertical dimension. These GAGAS standards include the concepts of reasonable assurance, significance, and audit risk for consideration when applying each of the four standards having to do with planning the audit; supervising staff; obtaining sufficient, appropriate evidence; and preparing audit documentation.

Table 3.2 Four Standards for Evaluation Fieldwork in Relation to Three Underlying, Pervasive Concepts

	Pervasive Concepts		
Standard	**Reasonable Assurance**	**Significance**	**Audit Risk**
Planning the Audit	Ensuring that the audit scope is reasonable (for example, in terms of audit objectives and realistic limitations on the project)	Developing a plan that adequately addresses all key issues that are material to the audit's validity (for example, making provisions for an adequate timeline and sufficient budget)	Providing safeguards to prevent or counteract suspected threats to the audit's integrity, such as an interested party that seeks to limit auditors' probing into certain relevant matters
Supervising Staff	Ensuring that the staff is sufficiently competent (for example, to meet audit objectives and earn credibility with audit audiences)	Making appropriate decisions in regard to the hiring of specialized consultants, for example	Obtaining legal counsel to help ensure the audit's legal viability, for example
Obtaining Sufficient, Appropriate Evidence	Ensuring that the full range of needed evidence will be obtained, for example	Achieving acceptable, documented levels of reliability and validity in obtained information, for example	Obtaining alternative forms of qualitative and quantitative evidence to assess and justify conclusions and recommendations, for example
Preparing Audit Documentation	Ensuring that reports will provide support for conclusions and recommendations with adequate supporting evidence, for example	Providing the level of documentation and validation of the documentation that audiences expect to see in the audit report, for example	Obtaining an independent, standards-based assessment of the draft audit report, for example

Source: Adapted from U.S. Government Accountability Office. (2007). *Government auditing standards* (GAO-07-731G). Washington, DC: U.S. Government Printing Office.

Reasonable Assurance

In addressing all four fieldwork standards, performance auditors need to do all they can to meet the GAGAS requirement of sufficient and appropriate evidence. Throughout the process of planning the audit, appointing and supervising staff, collecting evidence, and documenting and reporting on the audit, the auditors should develop a valid case for the level of confidence that users may confidently place in the audit's findings. Because performance audits vary in breadth of focus and availability of needed evidence, auditors need to exercise professional judgment in determining an appropriate scope for the audit and should inform users of the audit's strengths as well as its limitations in addressing audit objectives. Clearly, providing users of the audit with reasonable assurance of the audit's soundness for addressing its own objectives is a pervasive issue that should be systematically addressed throughout the process of planning, staffing, conducting, and reporting on an audit.

Significance

Significance concerns the relative importance of the various issues faced during the planning, staffing, conduct, and reporting of an audit within a defined context. Auditors need to carefully judge the significance of such audit matters as the type of audit to perform, the timeline for

the audit, the amount of required methodological rigor, the adequacy of the audit's budget, the sufficiency of audit staff qualifications, the provision of specialized training for audit staff members, the desirability of hiring specialized consultants, the need for legal experts to review plans and reports, the acceptability of documented levels of reliability and validity in obtained information, the breadth of audiences for the audit report, and the level of documentation required in the audit report. Because all audits are conducted under some level of constraints, auditors need to exercise professional judgment concerning a wide range of matters and their relative importance.

Audit Risk

GAGAS defines audit risk as

> the possibility that the auditors' findings, conclusions, recommendations, or assurance may be improper or incomplete, as a result of factors such as evidence that is not sufficient and/or appropriate, an inadequate audit process, or intentional omissions or misleading information due to misrepresentation or fraud. (U.S. Government Accountability Office, 2007, p. 123)

Throughout the process of planning, staffing, conducting, and reporting the audit, the auditors should be vigilant to identify and proactive to counteract such risks as overpromising what the audit can accomplish; agreeing to an overly restrictive time frame, budget, or set of rules for accessing records; depending unduly on the accuracy of the audited entity's database; errors in obtained information; and fraud or abuse in any aspect of the audit or the audited entity.

GAGAS notes:

> Audit risk can be reduced by taking actions such as increasing the scope of work; adding experts, additional reviewers, and other resources to the audit team; changing the methodology to obtain additional evidence, higher quality evidence, or alternative forms of corroborating evidence; or aligning the findings and conclusions to reflect the evidence obtained. (U.S. Government Accountability Office, 2007, p. 124)

Given the preceding principles to guide the fieldwork in performance audits, we now turn to the four fieldwork standards.

Fieldwork Standard 1: Planning

GAGAS provides extensive, detailed guidance for planning performance audits. Those who intend to use GAGAS to plan a performance audit or program evaluation are advised to reference and closely study pages 124 through 146. Here we summarize the main dimensions of that guidance.

The GAGAS standard for planning performance audits states, "Auditors must adequately plan and document the planning of the work necessary to address the audit objectives" (U.S. Government Accountability Office, 2007, p. 124). In planning an audit, auditors are directed

to assess significance and reduce audit risk to an appropriate level by exercising professional judgment in defining the audit's objectives, scope, and needed methods. Planning of these three elements is to be ongoing throughout the audit. Essentially, audit objectives denote the audit's subject matter and the performance questions to be answered based on evidence obtained and assessed against audit criteria. Scope refers to the audit's boundary and subject matter, and is directly tied to the audit objectives, with such specifications as the period of time to be reviewed, necessary documentation, and locations at which the fieldwork will be performed. The methodology comprises the methods and devices for gathering and analyzing the evidence needed to address the audit objectives, reduce audit risk to an acceptable level, and provide reasonable assurance that the evidence is sufficient and appropriate to support the auditor's findings and conclusions.

Staff development of the audit plan should take into account (1) the nature of the program; (2) the needs of potential users of the audit; (3) the relevance of internal control to the requirements of the audit; (4) the information system controls employed by the entity being audited; (5) pertinent legal and regulatory requirements, grants, and contracts; (6) risks of fraud or abuse having significance to audit objectives; (7) any relevant, ongoing investigations or legal proceedings; (8) results of pertinent previous audits; (9) the needed evaluative criteria; (10) the amount and type of required evidence and sources of relevant evidence; (11) the potential to use findings of other, similar audits; (12) required staff and other resources; and (13) needed arrangements for communicating with management, those charged with governance, and others.

Auditors are expected to prepare a written audit plan for each audit and to update the plan as necessary during the audit. GAGAS notes,

> The form and content of the written audit plan may vary among audits but should be sufficiently complete to enable the audit organization management to supervise audit planning. Using the plan they should be able to monitor and assess the audit's objectives; provisions for detecting and addressing risks; appropriateness of scope and methodology; sufficiency and appropriateness of evidence; collective competence of staff, supervisor, and specialists; and on-time performance. (U.S. Government Accountability Office, 2007, pp. 146–147)

Fieldwork Standard 2: Supervision

The standard states,

> Audit supervisors or those designated to supervise auditors must properly supervise audit staff. Supervisors are to guide and direct staff to ensure achievement of audit objectives, keep apprised of significant problems encountered, review work performed, and provide effective on-the-job training. (U.S. Government Accountability Office, 2007, p. 147)

Fieldwork Standard 3: Evidence

The standard states, "Auditors must obtain sufficient, appropriate evidence to provide a reasonable basis for their findings and conclusions." Under this fieldwork standard, "appropriateness is the measure of the quality of evidence that encompasses its relevance, validity, and

reliability in providing support for findings and conclusions related to the audit objectives," and "sufficiency is a measure of the quantity of evidence used to support the findings and conclusions related to the audit objectives" (U.S. Government Accountability Office, 2007, p. 147). In judging appropriateness and sufficiency of the evidence as a whole, auditors are expected to exercise professional judgment in determining whether enough sound evidence has been obtained to persuade a knowledgeable person that the findings are reasonable. In making such determinations, auditors are told to consider evidence to be unacceptably limited when its validity or reliability has not been assessed or cannot be assessed; tests have uncovered errors in the evidence; use of the evidence carries an unacceptable risk of leading to an incorrect or improper conclusion; or too little evidence has been obtained to justify the findings and conclusions. When the available evidence is judged inadequate, auditors are advised to take such further steps as seeking independent, corroborating evidence from other sources; redefining and possibly limiting audit objectives to what can be supported by the data; presenting the findings along with appropriate disclaimers and cautions; or reporting the cause of inadequate evidence, such as significant internal control deficiencies. In analyzing and interpreting obtained evidence, auditors may identify a situation related to the audit objectives, for example, a gap between the existing situation (condition) and the required or desired state (criteria), that may serve as the basis for recommending corrective action. To diagnose what problems need resolution, the auditors may then look for causes of the undesirable situation. Examples may include poorly designed policies, procedures, or criteria; staff with insufficient competence; inconsistent, incomplete, or incorrect implementation; deficiencies in internal control; or factors beyond the control of management. While steering auditors toward drawing cause-and-effect conclusions, GAGAS also notes that validly diagnosing causes of observed undesirable situations can be highly difficult to achieve, because deficient program performance often is a function of multiple complex factors.

Fieldwork Standard 4: Audit Documentation

According to this standard,

> Auditors must prepare audit documentation related to planning, conducting, and reporting for each audit. Audit documentation should contain sufficient detail to enable an experienced auditor, having no previous connection to the audit, to understand the nature, timing, extent, and results of audit procedures performed, the audit evidence obtained and its source, and the judgments and conclusions reached, including adequate supporting evidence. Auditors should document their support for findings, conclusions, and recommendations before issuing their audit report. (U.S. Government Accountability Office, 2007, pp. 156–157)

Pursuant to this fourth standard, auditors are advised to exercise professional judgment in documenting their audit work. Under GAGAS auditors should document the audit's objectives, scope, and methods; work performed to support significant judgments and conclusions, including transactions and records examined; and evidence of supervisory review of the

audit work prior to completing the final report. Auditors also are directed to document any departures from GAGAS requirements due to law, regulation, scope limitations, restrictions on accessing needed records, and so forth.

A continual process of documentation provides the principal support for the audit report, aids in conducting and supervising the audit work, and allows for review of audit quality. The documentation should be sufficiently detailed to make clear the audit's purpose, source, and conclusions, and it should be appropriately organized to clearly link the report's findings, conclusions, and recommendations. Audit organizations are expected to establish and implement reasonable policies for the proper storage and control of audit documentation for a time sufficient to satisfy legal and administrative requirements. Another intended use of audit documentation is to facilitate cooperation of government agencies in auditing programs of common interest so that auditors may use each other's work and avoid duplication of effort. Accordingly, auditors should, through such means as contracts, make appropriate individuals and audit documentation available in a timely manner to other, appropriate auditors or reviewers.

GAGAS Reporting Standards for Performance Audits

The three reporting standards for performance audits within GAGAS pertain to the form of reports, report contents, and report issuance and distribution. For each of these standards we give GAO's definition and then characterize GAO's elaboration of the standard.

Reporting Standard 1: Form

The standard states, "Auditors must issue audit reports communicating the results of each completed performance audit" (U.S. Government Accountability Office, 2007, p. 160). This first standard requires reports to be in a form that is appropriate for its intended use and in writing or otherwise retrievable—for example, a printed report, electronic media, letters, or briefing slides. Report forms should be chosen to communicate findings clearly to those charged with governance; appropriate officials of the audited entity; appropriate oversight officials; and, as applicable, the public. Reports should minimize misunderstanding of findings and facilitate follow-up to identify and assess corrective actions. If an audit is aborted, no report is issued. If, following issuance of a report, the auditors discover significant deficiencies in their findings and conclusions, the standard directs the auditors to so inform those charged with governance, the appropriate officials of the audited entity, and the appropriate officials of the organizations requiring or arranging for the audit, so that they will not continue to rely on findings or conclusions that are not supported by appropriate, sufficient evidence. Auditors should remove faulty reports from publicly accessible sources, such as Web sites, and post a public notification of such removal. Subsequently, the auditors should determine whether to recycle the audit process to obtain the needed valid evidence and issue a revised report containing appropriately supported findings or conclusions.

Reporting Standard 2: Report Contents

The standard states,

> Auditors should prepare audit reports that contain (1) the objectives, scope, and methodology of the audit; (2) the audit results, including findings, conclusions, and recommendations, as appropriate; (3) a statement about the auditors' compliance with GAGAS; (4) a summary of the views of responsible officials; and (5) if applicable, the nature of any confidential or sensitive information omitted. (U.S. Government Accountability Office, 2007, p. 161)

This standard directs auditors to define the audit objectives and audit criteria; identify the organizations, geographical locations, and time period covered; document the kinds and sources of evidence; and identify significant limitations or uncertainties concerning the auditors' assessment of the sufficiency and appropriateness of the overall set of evidence. Auditors are also expected to report any significant constraints on the audit work by information limitations or scope impairments, including denials of access to certain records or individuals. This standard also directs auditors to report deficiencies in internal control that are significant to audit objectives. In reporting the methodology they used, auditors are to identify significant assumptions underlying the audit, describe comparative techniques applied, and explain and justify any applied sampling design.

Auditors are also instructed to report all instances of fraud and illegal acts (unless they are inconsequential within the context of the audit objectives), significant violations of contract provisions or grant agreements, and significant abuse that have occurred or are likely to have occurred. Auditors are expected to report known or likely fraud, illegal acts, violations of contract provisions or grant agreements, or abuse directly to outside parties specified in law or regulation when management of the audited entity fails to satisfy legal or regulatory requirements to report such information to the specified external parties. Further, when entity management fails to take timely and appropriate steps to respond to known or likely fraud, illegal acts, violations of contract provisions or grant agreements, or abuse that is significant to the findings and conclusions and also involves direct or indirect funding from a government agency, auditors are expected to report such failure by the entity to the funding agency.

Conclusions are to be keyed to the audit objectives; supported by sufficient, appropriate evidence; and presented in the form of sound, logical inferences. Ideally, auditors will use sound conclusions to derive clear, convincing, concrete recommendations for action. Such recommendations for corrective action are seen as warranted if they are significantly keyed to audit objectives; supported by appropriate, sufficient evidence; and based on strong logic. Recommendations are defined as effective when they are addressed to parties that have the authority to act and when they are specific, practical, cost effective, and measureable.

This standard on report contents provides auditors with the following language for stating that their audit complied with GAGAS:

> We conducted this performance audit in accordance with generally accepted government auditing standards. Those standards require that we plan and perform the audit to

> obtain sufficient, appropriate evidence to provide a reasonable basis for our findings and conclusions based on our audit objectives. We believe that the evidence obtained provides a reasonable basis for our findings and conclusions based on our audit objectives. (U.S. Government Accountability Office, 2007, p. 169)

When auditors have not complied with all applicable GAGAS requirements, they are directed to include either a modified version of this compliance statement, including acknowledgment that not all standards were followed, or a straightforward statement indicating that the standards were not followed.

Auditors are advised to invite responsible officials of the audited entity and others to review and provide comments on a draft of the final report to help ensure that it is fair, complete, and objective. Auditors should include in the final report a copy or summary of any commentary on the audit report from the responsible officials of the audited program and any corrective actions the officials plan to pursue. Written commentary that can be reported verbatim is preferred, but if the auditors can only summarize oral commentary received, they should provide a copy of the summary to the respondents and ask them to verify its accuracy. As appropriate, the auditors should report their agreement or disagreement with the views or contemplated corrective actions presented by officials of the audited program, especially when these are inconsistent or in conflict with the draft report's findings, conclusions, or recommendations. Conversely, if respondents' criticisms of the draft report are found to be valid, the auditors are told to modify their report as necessary. If the audited entity declines the invitation to submit comments or does not respond in a reasonable period of time, the auditors should state in the report that the audited entity did not provide comments.

If certain pertinent information—that is sensitive, confidential, or prohibited from public disclosure—is omitted from the report, the audit report should identify the nature of the omitted information and state the reason or circumstances that make the omission necessary. This standard identifies a number of such reasons and provides recommendations for addressing these circumstances. Among the legitimate reasons for omitting certain information from a report are legal prohibitions, threats to public safety and security, and concern for the broader public interest. Suggestions for addressing these difficulties include communicating general information in a written report and communicating detailed information verbally or issuing a separate, classified or limited-use report containing the sensitive information only to persons who are authorized by law or regulation to receive it. Auditors may consult legal counsel about ways to legally address issues related to public records and other matters having to do with reporting information whose disclosure is problematic.

Reporting Standard 3: Report Distribution

This standard states, "Audit organizations in government entities should distribute audit reports to those charged with governance, to the appropriate officials of the audited entity, and to the appropriate oversight bodies or organizations requiring or arranging for the audits" (U.S. Government Accountability Office, 2007, p. 173).

As appropriate, auditors are advised to distribute copies of their reports to others authorized to receive them, such as officials with responsibility to act on audit findings and recommendations. When following GAGAS requirements in contracting to perform an audit, public accounting firms are instructed to clarify report distribution responsibilities with the engaging organization.

Internal auditors are advised to report results to parties who can ensure that the results are given due consideration. Before sending internal audit results to outside parties, internal auditors are instructed to assess potential risk to the organization; consult with senior management, legal counsel, or both; and control dissemination by identifying in the report those who are authorized to see and use it.

Supplemental Guidance

Offering supplemental guidance, Appendix 1 of GAGAS includes advice to assist auditors with implementing the standards. Although not establishing additional requirements, this guidance is intended to enhance users' understanding and application of the general, financial, attestation, and performance standards. Key areas of guidance include deficiencies in internal control; determining whether laws, regulations, or contract provisions or grant agreements are significant within the context of the audit objectives; laws, regulations, and guidelines that require the use of GAGAS; the role of those charged with governance in regard to accountability; management's role in regard to accountability; nonaudit services; the quality control system; types of evidence; appropriateness of evidence in relation to the audit objectives; and report quality elements.

Clearly, GAGAS is a valuable resource for program evaluators. In this chapter we have provided a summary of the parts of the document that are directly relevant to program evaluations. We strongly urge evaluators to obtain, study, and use this document plus updated editions as they are published. Program evaluators can strengthen their studies by regularly consulting this set of authoritative requirements, helpful explanations, and practical suggestions. They will find GAGAS useful in all stages of program evaluation work, including planning, contracting, staffing, conducting, reporting, and evaluating a program evaluation.

Using Evaluation Standards

Although the three sets of standards examined in this chapter vary in detail and substantive orientation, they are complementary, not contradictory. Fundamentally they are consistent in advocated principles but provide different emphases, cross-checks, levels of detail, and treatments of the requirements for sound evaluations. All three sets of standards are in substantial agreement as to what constitutes sound evaluation practices. Evaluations should be beyond reproach, with evaluators adhering to all relevant laws and ethical codes. Moreover, evaluators should produce valid findings and should be careful not to present unsupportable conclusions and recommendations. In addition, evaluators should carefully sort out their

roles as independent inquirers from their roles as social advocates and make sure that their evaluations are not corrupted by conflicts of interest. All three sets are grounded in the proposition that sound program auditing/evaluation is vital to the functioning of a healthy society. Service providers and governments must regularly subject their services to evaluation, and evaluators must deliver services that are legal, ethical, effective, safe, accounted for, and in the public interest. Standards are a powerful force for bringing about the needed sound evaluation services. Clearly, the three sets of standards together constitute a valuable resource offering principles, concepts, and procedures for evaluators and their clients.

Depending on particular evaluation assignments, the three sets may be used interchangeably or in concert. Comparisons of the substance of the 1994 Joint Committee program evaluation standards and the 1995 AEA guiding principles for evaluators (Shadish, Newman, Scheirer, & Wye, 1995b) have revealed key differences and similarities in the standards and principles (Covert, 1995; J. R. Sanders, 1995). Essentially everything covered by the AEA principles (Shadish et al., 1995b) was also covered by the Joint Committee's standards. However, the latter's coverage is broader and much more detailed, and it delves deeper into evaluation issues. No similar comparisons of the 2011 Joint Committee program evaluation standards; the government auditing standards (U.S. Government Accountability Office, 2007); and the 2004 AEA guiding principles for evaluators have been published. This would be a worthy project for a doctoral dissertation or other research investigation. In closing, we reprise each set of standards, discuss priorities for using each set, and outline a general process for applying standards.

The 2004 AEA guiding principles state that program evaluations should meet requirements for systematic inquiry, competence, integrity and honesty, respect for people, and responsibility for the general and public welfare. Of the three sets of standards, the guiding principles have the widest applicability and are the most general. They are officially endorsed by AEA and apply to program evaluations across a variety of government and social service sectors. They contain twenty-five important statements to support the five principles, but they lack detailed criteria and guidance. These standards arguably should be applied in all U.S. program evaluations, but due to their lack of specificity, they often function best as a secondary set of standards. Their applicability extends beyond the United States to all evaluators who decide to conduct their program evaluations in accordance with AEA's *Guiding Principles for Evaluators* (2004).

The 2011 Joint Committee program evaluation standards are focused on evaluations of educational programs in the United States and Canada. *The Program Evaluation Standards* stipulates that evaluations should meet requirements for utility, feasibility, propriety, accuracy, and evaluation accountability and provides extensive guidance and an assortment of illustrative cases. The development of the standards was sponsored by seventeen professional organizations concerned with improving education. Also, ANSI accredited these standards as the ones to be employed in evaluating educational programs in the United States.

The U.S. Government Accountability Office's *Government Auditing Standards* (2007) is focused on U.S. government–sponsored programs in all areas of government service. This document is grounded in ethical principles pertaining to the public interest; integrity; objectivity; proper use of government information, resources, and position; and professional behavior, and

provides general standards on independence, professional judgment, competence, and quality control and assurance. There are also specific standards for fieldwork in program audits that pertain to planning, supervision, obtaining sufficient, appropriate evidence, and audit documentation. And there are additional specific standards for reporting findings of program audits having to do with the form of reports, report contents, and report issuance and distribution. The general standards and the many specific topics in the chapters on performance audits are relevant to nongovernment as well as government evaluations in a wide range of program areas. Although they are intended for use in evaluating U.S. government programs, these standards have been used in countries across the world.

Evaluators can use a general nine-step process in applying all three sets of standards:

1. Become thoroughly familiar with each set of standards through systematic orientation and training.
2. Clarify the evaluation's purposes.
3. Clarify the evaluation's context.
4. Reach agreement with the client on which set or sets of standards will be applied and, if more than one set, which will be primary, secondary, or tertiary. In general, we suggest that the government auditing standards (U.S. Government Accountability Office, 2007) should be primary in evaluations of U.S. government programs, the program evaluation standards (Joint Committee, 2011) should be primary in evaluations of nongovernment educational programs in North America, and the AEA guiding principles (2004) should be primary in evaluations of nongovernment programs outside the field of education and secondary in all other program evaluations in the United States. However, recent experience has shown that some government agencies with a dominant interest in formative, improvement-oriented evaluation prefer to apply the Joint Committee program evaluation standards.
5. Orient and train stakeholders in the contents of the selected standards and their applicability to ensuring quality in the evaluation and ultimately assessing the program evaluation.
6. Apply the standards proactively through periodic checks on all aspects of the evaluation.
7. Give consideration to engaging an independent party to invoke the standards in conducting a formative or summative metaevaluation. Any formative application of the standards should include periodic written reports and feedback sessions aimed at strengthening the ongoing evaluation.
8. Apply the standards to assess the completed program evaluation. Such a summative metaevaluation will have more credibility if conducted by an independent evaluator.
9. Ensure that the summative metaevaluation report is released and effectively communicated to right-to-know audiences.

The primary tool for applying each set of standards is the full-length standards document, not merely a summary of the standards. In addition, checklists are available (www.wmich.edu/evalctr/checklists) to facilitate application of the Joint Committee program evaluation standards and the AEA guiding principles. GAO has issued a series of pamphlets

and maintains a useful Web site to help evaluators learn about and apply the most recent edition of *Government Auditing Standards* (www.gao.gov/govaud/ybk01.htm). All three standards documents emphasize that the standards are general guidelines, and that evaluators and their clients should consult and employ much more specific material when dealing with such details as design, measurement, case studies, statistics, reporting, and contracting.

Summary

Professionally defined standards are essential to the sound, effective practice of evaluation. They provide authoritative criteria for guiding and judging evaluations and, when met, militate against bias and enhance credibility. To be authoritative and credible, evaluation standards must reflect a general consensus by experts in the conduct and use of evaluation. The past three decades have seen three organized, U.S.-based efforts to develop, evolve, and apply evaluation standards. The three sets of standards discussed in this chapter have different foci and emphases but in general are compatible and complementary. Fundamentally, all three sets present general, appropriately vetted and endorsed principles, each exemplifying the essential meaning of a standard.

Since 1975 the Joint Committee, under the sponsorship of about fifteen professional societies in the United States and Canada, has been issuing and updating standards for evaluations of programs, personnel, and students. The committee's standards, in all three areas, call for evaluations to be useful, feasible, proper, and accurate, and, in the 2011 edition of *The Program Evaluation Standards,* also accountable. Each of approximately thirty standards in each set is elaborated with specific guidelines, pitfalls to avoid, and one or two illustrative cases. The Joint Committee's standards are accredited by ANSI and designed for application to educational evaluations that occur in the United States and Canada.

Since 1992 AEA has been developing general guiding principles for evaluators. In the 2004 standards document, twenty-five briefly stated guidelines are grouped into the categories of systematic inquiry, competence, integrity/honesty, respect for people, and responsibilities for general and public welfare. These guidelines are aimed at program evaluations across disciplines and intended for use by evaluators, anywhere, who choose to meet AEA's requirements for sound evaluations.

In 1972 GAO began issuing government auditing standards, which until the issuance of the 2002 set of standards dealt almost exclusively with the financial aspects of federal programs. The 2007 edition includes program audits as one of the foci of general standards and presents chapters containing fieldwork standards and reporting standards for performance audits. The general standards have to do with (1) the independence of the audit organization and its individual auditors; (2) the exercise of sound professional judgment in conducting and reporting audits, exercising quality control, and engaging external peer reviews; (3) the competence and continuing education of audit staff; and (4) provisions for quality control to provide reasonable assurance of compliance with applicable auditing standards. These general standards are considered mandatory when performing audits requiring application of the government auditing standards. Especially noteworthy in the 2003 and 2007 editions is the section

on independence, which prohibits auditors from simultaneously providing the same entity with both auditing and consulting services, because such commingling of services is seen as an unacceptable conflict of interest. Labels for the fieldwork standards for performance audits are Planning, Supervision, Evidence, and Audit Documentation. The three reporting standards for performance audits pertain to the form of reports, report contents, and report issuance and distribution.

The concluding section of this chapter provided a general process for applying all three sets of standards. It called for studying the actual standards documents, clarifying the subject evaluation's purpose and context, deciding which set or sets of standards to apply, training stakeholders in the requirements of the selected standards, applying the standards to all aspects of the evaluation and throughout the evaluation process, deciding whether to obtain an independent metaevaluation based on the standards, and obtaining and disseminating the findings of a final, summative metaevaluation based on the standards.

REVIEW QUESTIONS

1. State reasons why adhering to standards for program evaluation is in the interest of both program staff and program recipients.
2. For each of the three sets of standards, explain why evaluators should feel confident that the standards presented in this chapter provide authoritative direction for guiding and assessing evaluations.
3. Respond to the claim that the introduction of standards, with their objective criteria, has strengthened the professionalism of all evaluators.
4. Identify eight to ten specific functions of standards for program evaluation.
5. Outline the most important features of each of the three sets of standards presented in this chapter.
6. Provide a list of dangers inherent in not closely referencing standards during the course of an evaluation, and cite cases in which these dangers caused an evaluation to fail or be discredited.
7. Examine and comment on the claim that the Joint Committee program evaluation standards are relevant and useful when planning and conducting metaevaluations.
8. Comment on the assertion that AEA's *Guiding Principles for Evaluators* offers general codes of behavior supported by normative statements; does not guide an evaluator in a direct, operational sense; but does have ethical utility.
9. List ten or more significant ways in which following the government auditing standards ensures the accuracy and credibility of audit results.
10. This chapter has stated that all three sets of standards may be applied to a study in concert, individually, or interchangeably. Briefly outline an evaluation situation for each of these types of application.

Group Exercises

Exercise 1

Contrast the evaluation field without published standards (in other words, the field prior to the 1980s) with the situation today. Discuss the salient differences from the point of view of both evaluator and client. In your discussion, reach conclusions about the influence of published standards on the evaluation field's progress toward attaining the stature of a mature, highly respected profession.

Exercise 2

Prior to addressing this exercise, ask one group member to obtain and provide other members with copies of a completed evaluation report (which should not be too extensive). After group members have studied the report, use this chapter's summary of the Joint Committee's 2011 program evaluation standards to evaluate the evaluation (that is, to conduct a metaevaluation). Reach and itemize conclusions about the evaluation's strengths and weaknesses.

Exercise 3

Discuss the kind of situation in which an evaluator would predominantly use, at least as a primary tool, each of these documents:

- *The Program Evaluation Standards*
- *Guiding Principles for Evaluators*
- *Government Auditing Standards*

Exercise 4

The managing director of a large manufacturing firm has a problem: his section leaders have reported to him that an evaluation of a new program in the firm has caused growing anxiety among the workforce, principally because a whole range of established workers' rights will be abrogated by the evaluation. She enlists the services of an experienced evaluator to evaluate the program evaluation. Which of the Joint Committee's standards would be especially relevant to the workers' concerns about rights violations, and why?

Exercise 5

In the course of conducting a series of workshops to help an organization's technical support staff learn about and make better use of computer-assisted design and other computer technology, the contracting instructor makes it known that, beyond offering technology instruction, he is also an expert in evaluating technology services. Because the organization's head judges her organization's use of technology to be deficient, she invites the workshop provider to contract both for advising on how to strengthen the organization's use of technology and for evaluating the organization's technology capabilities and performance. Based on what

you have read about professional standards for evaluations, how should the workshop provider respond to this request? Which set of standards is most relevant to this situation? What particular standards would provide the best basis for formulating a sound, professionally defensible response, and why? What might be an inappropriate response, and what possible negative consequences could accompany such a response?

Notes

1. The Sarbanes-Oxley Act of 2002 is a federal law that set new or enhanced standards for all U.S. public company boards and management and public accounting firms. It was enacted as a reaction to major corporate and accounting scandals, including those affecting Enron, Tyco International, Adelphia, Peregrine Systems, and WorldCom. These scandals cost investors billions of dollars when the share prices of affected companies collapsed and shook public confidence in the nation's securities markets. The act assigns additional corporate board responsibilities, stipulates criminal penalties for violations of the act, and requires the Securities and Exchange Commission to implement rulings on compliance with the law. The act created a new, quasi-public agency, the Public Company Accounting Oversight Board, charged with overseeing, regulating, inspecting, and disciplining accounting firms in their roles as auditors of public companies. The act also covers such issues as auditor independence, corporate governance, internal control assessment, and enhanced financial disclosure.
2. The initial committee had the following representatives of user groups: William Ellena, Homer Elseroad, Philip Hosford, William Mays Jr., Bernard McKenna, James Mecklenburger, and James Ward. It had the following representatives of methodological specialties: Henry Brickell, Donald Campbell, Ronald Carver, Esther Diamond, Egon Guba, Robert Linn, George Madaus, Wendell Rivers, Lorrie Shepard, and Daniel Stufflebeam (chair).
3. The sponsors of the Joint Committee, as of publication of *The Program Evaluation Standards* in 2011, were the American Association of School Administrators, American Counseling Association, American Educational Research Association, American Evaluation Association, American Indian Higher Education Consortium, American Psychological Association, American Society for Curriculum Development, Canadian Evaluation Society, Canadian Society for the Study of Education, Consortium for Research and Educational Accountability and Teacher Evaluation, Council of Chief State School Officers, National Association of Elementary School Principals, National Association of School Psychologists, National Council on Measurement in Education, National Education Association, National Legislative Program Evaluation Society, and National Rural Education Association.
4. The term *attestation engagement* is not commonly seen in the program evaluation literature. In such an engagement, an auditor issues an examination, a review, or an agreed-on procedural report on a subject matter or an assertion about a subject matter, pursuant to criteria selected by another party. Attestation engagements can cover a broad range of financial or nonfinancial objectives and result in various types of opinions depending on the user's needs.

Suggested Supplemental Readings

Covert, R. W. (1995). A twenty-year veteran's reflections on the guiding principles for evaluators. In W. R. Shadish, D. L. Newman, M. A. Scheirer, & C. Wye (Eds.), *Guiding principles for evaluators* (pp. 35–45). New Directions for Program Evaluation, no. 66. San Francisco, CA: Jossey-Bass.

Evaluation Research Society Standards Committee. (1982). Evaluation Research Society standards for program evaluation. In P. H. Rossi (Ed.), *Standards for evaluation practice* (pp. 7–19). New Directions for Program Evaluation, no. 15. San Francisco, CA: Jossey-Bass.

Joint Committee on Standards for Educational Evaluation. (1981). *Standards for evaluations of educational programs, projects, and materials*. New York, NY: McGraw-Hill.

Joint Committee on Standards for Educational Evaluation. (1988). *The personnel evaluation standards*. Thousand Oaks, CA: Corwin Press.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Joint Committee on Standards for Educational Evaluation. (2003). *The student evaluation standards*. Thousand Oaks, CA: Corwin Press.

Joint Committee on Standards for Educational Evaluation. (2009). *The personnel evaluation standards: How to assess systems for evaluating educators* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Joint Committee on Standards for Educational Evaluation. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Sanders, J. R. (1995). Standards and principles. In W. R. Shadish, D. L. Newman, M. A. Scheirer, & C. Wye (Eds.), *Guiding principles for evaluators* (pp. 47–53). New Directions for Program Evaluation, no. 66. San Francisco, CA: Jossey-Bass.

Shadish, W. R., Newman, D. L., Scheirer, M. A., & Wye, C. (1995). Developing the guiding principles. In W. R. Shadish, D. L. Newman, M. A. Scheirer, & C. Wye (Eds.), *Guiding principles for evaluators* (pp. 3–18). New Directions for Program Evaluation, no. 66. San Francisco, CA: Jossey-Bass.

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, CA: Jossey-Bass.

U.S. General Accounting Office. (2002). *Government auditing standards: Amendment no. 3 Independence* (GAO-02-388G). Washington, DC: U.S. Government Printing Office.

U.S. Government Accountability Office. (2003). *Government auditing standards* (GAO-03-763G). Washington, DC: U.S. Government Printing Office.

U.S. Government Accountability Office. (2007). *Government auditing standards* (GAO-07-731G). Washington, DC: U.S. Government Printing Office.

PART TWO

AN EVALUATION OF EVALUATION APPROACHES AND MODELS

The seven chapters in Part Two identify and assess approaches often employed to evaluate programs. Chapter 4 provides background information for reviewing evaluation approaches. Chapters 5 through 9 characterize and assess pseudoevaluation, quasi-evaluation, improvement- and accountability-oriented, social agenda and advocacy, and eclectic approaches, respectively. Chapter 10 provides a consumer report evaluation of nine of the highest-rated or most used approaches.

CHAPTER 4

BACKGROUND FOR ASSESSING EVALUATION APPROACHES

The chapters in Part One have helped provide a firm understanding of the basic concepts and principles of program evaluation. The chapters in Part Two identify and assess twenty-three approaches often employed to evaluate programs. The evaluation approaches reviewed here are in varying degrees unique and cover most program evaluation efforts. Our objective is to help readers decide which of the approaches are most worthy of application and further development, and which are best abandoned.

The approaches reviewed here emerged mainly in the United States between 1960 and 2000. Six of the approaches, labeled "pseudoevaluations," reflect the political realities of evaluation and are often used illegitimately to falsely characterize (or hide) a program's value. Pseudoevaluations have often arisen from expediency, without due consideration given to the ethics or professional soundness of their design and implementation. Unfortunately, repeated use of these approaches has all too often given them a veneer of respectability and legitimacy.

We review these approaches in the hope of helping evaluators and clients identify, avoid, or expose misleading or blatantly corrupt studies offered in the name of evaluation. The remaining seventeen approaches are typically used legitimately to judge programs. They are divided into eight quasi-evaluations (approaches narrowly focused on answering one or a few questions or using mainly one method), three improvement- and accountability-oriented approaches (aimed at determining an evaluand's merit and worth), four social agenda and advocacy approaches (usually dedicated to righting social injustices), and two

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **The system used in this book to organize and classify twenty-three evaluation approaches, as well as alternative systems for doing so**
- **A historical overview of proposed evaluation approaches and analyses of such approaches**
- **The definitions of five categories of evaluation approaches: pseudoevaluation, quasi-evaluation, improvement oriented and accountability oriented, social agenda and advocacy, and eclectic**
- **Nine descriptors used in this book to characterize selected evaluation approaches**

eclectic approaches (drawing selectively from all available evaluation concepts and methods to serve the needs of a particular user group). We have characterized each approach; assessed its strengths and weaknesses; and considered when and how it is best applied (if it should be applied at all). All legitimate approaches are enhanced when keyed to professional standards for evaluations, and, accordingly, we have assessed the most promising approaches against the Joint Committee on Standards for Educational Evaluation's *Program Evaluation Standards* (2011).

The development of many of the reviewed approaches was spurred by a number of seminal writings. These include, in chronological order, publications by Flexner (1910); R. W. Tyler (1932, 1942, 1950); Fisher (1951); Lindquist (1953); Campbell and Stanley (1963); Cronbach (1963); Kaplan (1964); Stufflebeam (1966b); R. W. Tyler (1966); Stufflebeam (1967); B. G. Glaser and Strauss (1967); Metfessel and Michael (1967); Scriven (1967); Stake (1967); Suchman (1967); Alkin (1969); Guba (1969); Provus (1969); Lessinger (1970); Stufflebeam et al. (1971); Hammond (1972); Parlett and Hamilton (1972); Weiss (1972); House (1973); Rippey (1973); Eisner (1975); Glass (1975); Wolf (1975); Cook and Reichardt (1979); Cook and Campbell (1979); Cronbach and Associates (1980); House (1980); Patton (1980); the Joint Committee (1981); H. M. Levin (1983); Stake (1983); Bickman and Peterson (1990); Chen (1990); W. L. Sanders and Horn (1994); Fournier (1995); Pawson and Tilley (1997); Henry, Julnes, and Mark (1998); Shadish, Cook, and Campbell (2002); Cousins (2003); Brinkerhoff (2003); and Mertens (2009). These publications and those of other authors and scholars offered noteworthy alternative approaches to program evaluation.

Over the years, a rich literature on a wide variety of alternative program evaluation approaches developed. See, for example, Alkin (2004); Boruch (1994, 2003); Campbell (1988); Chelimsky (1987); T. D. Cook and Reichardt (1979); Cousins (2003); Cousins and Earl (1992); Cronbach (1982); Davis and Salasin (1975); Denny (1978); Eisner (1983); Fetterman (1984, 1994); Fitzpatrick, Sanders, & Worthen (2011); Flinders and Eisner (2000); Greene (1988a); Guba (1978); Guba and Lincoln (1981, 1989); Henry, Julnes, and Mark (1998); Hofstetter and Alkin (2003); House and Howe (1998, 2000a, 2000c, 2003); Joint Committee (1994); Karlsson (1998); Kee (1995); Kellaghan and Stufflebeam (2003); Kidder and Fine (1987); Kirst (1990); Koretz (1996a, 1996b); Levin (1983); Levine (1974); Lincoln and Guba (1985); Linn, Baker, and Dunbar (1991); MacDonald (1975); Madaus, Scriven, and Stufflebeam (1983); Madaus and Stufflebeam (1988); Mathison (2005a); Mehrens (1972); Messick (1994); National Science Foundation (1997); Nave, Miech, and Mosteller (2000); Nevo (1993); Owens (1973); Patton (1982, 1990, 1994, 1997, 2000, 2003, 2008, 2010); Platt (1992); Popham (1969); Popham and Carlson (1977); Provus (1971); Rogers (2000); Rossi, Lipsey, Freeman, and Rosenbaum (in various configurations; 1979, 1999, 2004); J. R. Sanders (1992); W. L. Sanders (1989); Schwandt (1984, 1989); Scriven (1991, 1993, 1994a, 1994b, 1994c); Shadish, Cook, and Leviton (1991); M. F. Smith (1986, 1989); N. L. Smith (1987); Stake (1975b, 1986, 1988, 1995); Stufflebeam (1997, 2001b); Stufflebeam, Madaus, and Kellaghan (2000); Stufflebeam and Shinkfield (1985); Torres (1991); Tsang (1997); Tymms (1995); Webster (1975, 1995); Webster, Mendro, and Almaguer (1994); Weiss (1995); Whitmore (1998); Wholey (1995); Worthen and Sanders (1987); Worthen, Sanders, and Fitzpatrick (1997); and Yin (1992, 2009).

Following a period of relative inactivity in the 1950s, a succession of international and national forces stimulated the expansion and development of evaluation theory and practice. The main influences were the efforts to vastly strengthen the U.S. defense system spawned by the Soviet Union's 1957 launch of Sputnik I; the new U.S. laws in the 1960s to serve minorities and persons with disabilities equitably; federal evaluation requirements of the Great Society programs initiated in 1965; the movement begun in the 1970s to hold educational and social organizations accountable in regard to both prudent use of resources and achievement of objectives; the stress on excellence in the 1980s as a means of increasing the international competitiveness of the United States; and the increasing trend in the 1990s and beyond for various organizations, both inside and outside the United States and across disciplines, to employ evaluation to ensure quality, competitiveness, and equity in delivering services. In pursuing reforms, American society has repeatedly pressed schools and colleges, health care organizations, government organizations, manufacturers, and various social welfare enterprises to show through evaluation whether services and improvement efforts have been succeeding.

Evaluation Approaches

This book uses the term e*valuation approach* along with *evaluation model* because the former is broad enough to cover illicit as well as laudatory practices. Also, beyond covering both creditable and noncreditable approaches, some authors of evaluation approaches say that the term *model* is too demanding and restrictive to cover their published ideas about how to conduct program evaluations. Moreover, some leading proponents of program evaluation see their work as evolutionary, and therefore some flexibility about aspects of their approaches is required. But for these two considerations, the term *model* would have been used to encompass most of the evaluation proposals discussed in this book. This is so because most of the presented approaches are idealized or model views for conducting program evaluations according to their authors' beliefs and experiences.

Importance of Studying Alternative Evaluation Approaches

The study of alternative evaluation approaches is important for professionalizing program evaluation, which will lead to its scientific operation and advancement. Careful, professional study of alternative ways of conducting program evaluations will help evaluators discredit approaches that violate sound principles of evaluation and legitimize and strengthen those that follow the principles. Scientifically, such reviewing of approaches will help evaluation researchers identify, examine, and address conceptual and technical issues pertaining to the development of the evaluation discipline. Operationally, taking a critical view of alternatives can help evaluators consider, assess, and selectively apply optional and appropriate evaluation frameworks. Such review will also provide a sound basis for evaluation training. The main value in studying alternative program evaluation approaches lies in discovering their strengths and weaknesses and determining the circumstances under which each is appropriately applied. Such analysis will help determine which ones merit substantial use, determine when and how they

are best applied, obtain direction for improving the approaches and devising better alternatives, and strengthen one's ability to conceptualize hybrid approaches to program evaluation.

The Nature of Program Evaluation

We take a broad view of program evaluation, seeing it as encompassing assessments of any coordinated set of activities directed at achieving goals. Program evaluations may be conducted in business and manufacturing enterprises (both large and small); community or state organizations; charitable foundations; local, state, and federal government agencies; welfare and voluntary groups; or any other entities where activities have been discernibly planned to meet assessed needs and defined goals. More specific examples of program evaluations are assessments of ongoing, cyclical programs, such as those having to do with school curricula, food stamps, housing for the homeless, and annual influenza inoculations; of time-restricted projects, such as development and dissemination of a fire prevention guide and development of a new instrument for evaluating the performance of factory workers; and of national, regional, or state systems of services, such as those provided by regional educational service organizations and a state's department of natural resources. Program evaluations overlap with and yet are distinguishable from other forms of evaluation, especially evaluations of students, personnel, policies, and products, among others.

Previous Classifications of Alternative Evaluation Approaches

In analyzing the twenty-three evaluation approaches, we considered prior assessments of program evaluation's state of the art. Stake's analysis (1974) of nine program evaluation approaches provided a useful application of advance organizers (his notion of cues that evaluators use to set up a study). Hastings's review (1976) of the growth of evaluation theory and practice helped put the evaluation field in historical perspective. Guba's book *The Paradigm Dialog* (1990) and his 1977 presentation and assessment of six major philosophies in evaluation were provocative. House's analysis (1983) of approaches illuminated important philosophical and theoretical distinctions. Scriven's writings (1991, 1994a) on evaluation as a transdiscipline and the transdiscipline of evaluation helped us sort out different evaluation approaches; they were also invaluable in helping us see the approaches in the broader context of evaluations focused on various objects other than programs. The books *Evaluation Models: Viewpoints on Educational and Social Services Evaluation* (Madaus et al., 1983; Stufflebeam et al., 2000); *Evaluation Models* (Stufflebeam, 2001b); and *International Handbook of Educational Evaluation* (Kellaghan & Stufflebeam, 2003) provided previous inventories and analyses of evaluation models. All of the assessments helped sharpen the issues addressed.

Classification and Analysis of the Twenty-Three Evaluation Approaches

In characterizing and assessing evaluation approaches, we have classified the various kinds of activities conducted in the name of program evaluation on the basis of their level of conformity to the definition of evaluation given in the 1994 edition of the Joint Committee's

Program Evaluation Standards. According to that definition, evaluation is the assessment of something's worth or merit. This definition should be widely acceptable because it is consistent with common dictionary definitions of evaluation; because the Joint Committee (1981, 1988, 1994, 2003, 2009, 2011) used it in developing professional standards for evaluations of programs, personnel, and students; and because the Joint Committee's standards are accredited by the American National Standards Institute. In Chapter 5 it will become apparent that many studies done in the name of program evaluation either do not conform to the essential meaning of evaluation or directly oppose it.

Using the definition of evaluation just given, we classified program evaluation approaches in consideration of the extent to which they focus mainly, somewhat, or not at all on judging a program's value. Accordingly, we identified five categories of evaluation approaches. The first category includes approaches that promote invalid or incomplete findings (referred to as pseudoevaluations), and the other four include approaches that agree, more or less, with the Joint Committee's definition of evaluation (quasi-evaluation, improvement- and accountability-oriented, social agenda and advocacy, and eclectic approaches). Of the twenty-three program evaluation approaches that are described, six are classified as pseudoevaluations, eight as quasi-evaluations, three as improvement- and accountability-oriented approaches, four as social agenda and advocacy approaches, and two as eclectic approaches. Each approach is characterized in terms of nine descriptors: (1) advance organizers—that is, the main cues that evaluators use to set up a study; (2) main purposes served; (3) sources of questions addressed; (4) questions that are characteristic of the approach; (5) methods typically employed; (6) pioneers in conceptualizing the approach plus others who have extended its development and use; (7) key considerations in determining when to use the approach; (8) strengths of the approach; and (9) weaknesses of the approach.

Nine approaches that appeared most worthy were then selected for a consumer report analysis, which appears in Chapter 10. We evaluated these approaches against the requirements of the Joint Committee's 2011 program evaluation standards to obtain judgments—of poor, fair, good, very good, or excellent—of each approach's utility, feasibility, propriety, accuracy, evaluation accountability, and overall merit. The judgments of each of the nine approaches were reached using a specially prepared checklist. For each of the thirty program evaluation standards, the checklist contained checkpoints representing the standard's key requirements. We rated each of the evaluation approaches on each of the thirty Joint Committee program evaluation standards by judging whether the approach, as defined in the literature and otherwise known, satisfactorily addressed each of the checkpoints. We rated the approaches based on our knowledge of the Joint Committee's thirty standards, our many years of studying the various evaluation approaches, our experience in seeing and assessing how some of these models and approaches worked in practice, and our personal experiences in working with authors or leading proponents of all nine approaches.

One of us (the first author) chaired the Joint Committee during its first thirteen years and led the development of the first editions of the documents containing the program and personnel evaluation standards. And both of us have collaborated with many of the

evaluation field's leading developers of evaluation approaches. We have been privileged to collaborate with Robert Brinkerhoff (principal developer of the Success Case Method), Donald Campbell (a leading developer of experimental and quasi-experimental design), J. Bradley Cousins (a leading figure in participatory evaluation), Lee Cronbach (a leading developer of decision-oriented evaluation), Elliot Eisner (developer of the connoisseurship and criticism approach), Gene Glass (leading developer of meta-analysis), Egon Guba (principal developer of constructivist evaluation), Ernest House (leading developer of the deliberative democratic approach), Michael Patton (developer of utilization-focused evaluation), Michael Scriven (leading developer of consumer-oriented program evaluation), Robert Stake (author of responsive evaluation and a leading proponent of case study evaluation), Ralph W. Tyler ("father of educational evaluation" and principal developer of objectives-based evaluation), and William Sanders (author of value-added evaluation). We feel indeed privileged to have known and collaborated with these figures. Clearly they are giants in the realm of developing systematic approaches to evaluation. Our experiences in working with them have greatly enriched the perspective from which we have written this book. Accordingly, we owe all of them a profound debt of gratitude.

Caveats

We acknowledge, without apology, that the assessments of the approaches in this part of the book are based on our best judgments. We have taken no poll, and no definitive research exists, to represent a consensus on the characteristics, strengths and weaknesses, and comparative merits of the different approaches. We also acknowledge a conflict of interest, because the first author developed one of the rated approaches, the context, input, process, and product (CIPP) model discussed in Chapter 7. Our main means of addressing this conflict are to acknowledge it and to make our assessment criteria and procedures explicit, so that others can assess them and, if they choose, use them to conduct their own analyses, render their own judgments of the approaches, and then compare their conclusions with ours.

Our analyses reflect a combined total of fifty years of experience in applying and studying different evaluation approaches. In a sense, with our relevant backgrounds and experience, we are the instruments employed in this analysis. Perhaps our assessments fall best under the category of Eisner's connoisseurship and criticism approach. We hope our analyses will be useful to evaluators, evaluation clients, and evaluation students in selecting and applying evaluation approaches. Moreover, we hope evaluation researchers will find our analyses useful for generating and testing hypotheses concerning the relative effectiveness of the different evaluation approaches and also for developing better approaches.

Finally, we have mainly looked at the approaches as relatively discrete ways to conduct evaluations. In reality, there are many occasions when it is functional to mix and match different approaches, as seen in the realm of eclectic evaluation. A careful analysis of such combinatorial applications no doubt would produce several additional hybrid approaches that might merit examination. That analysis is beyond the scope of this book.

Summary

Systematic examinations of evaluation approaches are increasingly important for professionalizing program evaluation. This is especially so given the existence of numerous illegitimate approaches, which can sometimes harm evaluation as a discipline and professional field of practice. The primary rationale for studying alternative program evaluation approaches is to determine their strengths and weaknesses and the circumstances under which each might be appropriate and useful.

Evaluation approaches can be, and have been, classified and described in numerous ways. In this chapter we have described a system of classification that includes pseudoevaluations, quasi-evaluations oriented to a narrow set of questions or a single method, improvement- and accountability-oriented approaches, social agenda and advocacy approaches, and eclectic evaluation approaches. In addition, each approach, to be discussed in later chapters, has been characterized in terms of nine descriptors: (1) advance organizers—that is, the main cues that evaluators use to set up a study; (2) main purposes served; (3) sources of questions addressed; (4) questions that are characteristic of the approach; (5) methods typically employed; (6) pioneers in conceptualizing the approach plus others who have extended its development and use; (7) key considerations in determining when to use the approach; (8) strengths of the approach; and (9) weaknesses of the approach.

REVIEW QUESTIONS

1. Summarize the rationale we have given for classifying evaluation approaches into discrete categories.
2. Identify and define each of the five categories of approaches discussed in this book.
3. Rank-order the five categories based on your perception of their relative utility for you in designing and conducting evaluations or in using their findings; then write a justification for the rank you gave to each category.
4. State the definition of evaluation we have used to classify and assess different evaluation approaches; then summarize the reasons given for using this definition. Finally, write your opinion as to whether this or some other definition is an appropriate basis for classifying and judging evaluation approaches.
5. List the nine descriptors selected to characterize evaluation approaches. Then give your assessment of the extent to which these descriptors are appropriate and sufficient for characterizing and judging evaluation approaches.
6. Summarize (a) our acknowledgment of a conflict of interest in our evaluation of evaluation approaches and (b) how we addressed this conflict. Then (c) give your assessment of our

treatment of this conflict, and (d) state whether, and if so how, we should have addressed the conflict differently.

7. Summarize and give your assessment of our stated plan for systematically characterizing and evaluating the twenty-three selected evaluation approaches. Then describe an approach you could use to independently evaluate the twenty-three approaches.

Group Exercise

For this exercise, group members should select a completed evaluation, study its final report, analyze the evaluation according to the analytical approach presented in this chapter, have a preliminary discussion to solidify understanding of the chapter, and have a further discussion to assess the chapter's criteria and procedures for classifying and characterizing evaluation approaches. Here are the assigned steps for completing the exercise:

1. Select a completed evaluation and read its final report.
2. Based on this book's scheme for classifying evaluation approaches, determine whether your group's selected evaluation is best classified as a pseudoevaluation, quasi-evaluation, improvement- and accountability-oriented evaluation, social agenda or advocacy evaluation, or eclectic evaluation.
3. Review and justify your classification of the evaluation.
4. Make a list of the evaluation's salient characteristics, using the nine descriptors provided in this chapter.
5. As needed, review chapter contents to update your understanding of the chapter's approach to classifying and characterizing evaluation approaches.

Suggested Supplemental Readings

Alkin, M. C. (Ed.). (2004). *Evaluation roots: Tracing theorists' views and influences*. Thousand Oaks, CA: Sage.

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2010). *Program evaluation: Alternative approaches and practical guidelines* (4th ed.). Upper Saddle River, NJ: Pearson.

House, E. R. (1983). Assumptions underlying evaluation models. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 45–64). Norwell, MA: Kluwer.

Schwandt, T. A. (1984). *An examination of alternative models for socio-behavioral inquiry*. Unpublished doctoral dissertation, Indiana University, Bloomington.

Scriven, M. (1994b). Evaluation as a discipline. *Studies in Educational Evaluation*, *20*, 147–166.

Stake, R. E. (1974). *Nine approaches to educational evaluation*. Urbana: University of Illinois, Center for Instructional Research and Curriculum Evaluation.

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

CHAPTER 5

PSEUDOEVALUATIONS

Background and Introduction

Because this book is aimed at examining and explaining the state of the art in evaluation, it is necessary to discuss bad and questionable practices as well as best efforts. Evaluators and their clients are sometimes tempted to shade, selectively release, overgeneralize, or even falsify findings. In addition, evaluators might falsely characterize constructive efforts—such as providing evaluation training or developing an organization's evaluation capability—as sound evaluation. Or they might unwittingly conduct an evaluation that serves a hidden, corrupt purpose. Others—lacking true knowledge of evaluation planning, procedures, and standards—may feign evaluation expertise while producing and reporting false outcomes. In addition, commercial enterprises that obtain and publish, on their Web sites, customer feedback on purchased products or services might or might not be providing representative, valid assessments. Although such activities conducted in the name of evaluation might look like sound evaluations, they are aptly termed "pseudoevaluations" if they fail to produce valid assessments of merit or worth, and if evaluators do not report these assessments to all right-to-know audiences.

Pseudoevaluations often are motivated by political objectives or profit motives. For example, persons holding or seeking authority may present unwarranted claims about their achievements or the faults of their opponents, or they may hide potentially damaging information. Or a "do-gooder" evaluator wanting to pacify, secure acceptance from, or improve the evaluation capabilities of a

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **Public relations studies**
- **Politically controlled studies**
- **Pandering evaluations**
- **Evaluation by pretext**
- **Empowerment under the guise of evaluation**
- **Customer feedback evaluation**

group of unwary, unsophisticated evaluees or other stakeholders may compromise an evaluation's independence and water down results to win the confidence of evaluees or help them gain power. Corrupted evaluations are considered here because they deceive through supposedly valid studies and can be used by those in power to mislead constituents or to gain and maintain an unfair advantage over others, especially those with little power. Also, pseudoevaluations are considered because they threaten the integrity of the evaluation profession. Conversely, consistent and best practice, widely understood by both evaluator and client, will elevate the profession.

We have identified six pseudoevaluation approaches for discussion in this chapter: public relations studies; politically controlled studies; pandering evaluations; evaluation by pretext; empowerment under the guise of evaluation; and selective, possibly biased release of customer feedback. They are primarily distinguished through flaws in regard to truth seeking, representativeness of findings, writing and editing of reports, and dissemination of findings. The main objective of evaluators conducting public relations studies is not to seek truth but instead to acquire and broadcast information that provides a favorable, though often false, impression of a program. Evaluators undertaking politically controlled studies seek truth but inappropriately control the release of findings to right-to-know audiences. In pandering evaluations, evaluators tell clients what they want to hear rather than what is true; they do so to obtain favors from clients, including future evaluation contracts. In an evaluation by pretext, the evaluator begins with a preferred conclusion or decision and rigs data to support a predetermined outcome. In empowerment under the guise of evaluation, an external evaluator pursues the laudable objective of helping the client group develop evaluation expertise and mainstream evaluation in an organization, but in so doing gives the client group the authority to write, rewrite, edit, or selectively release that external evaluator's so-called independent report. In such evaluation work, the desirable end of helping clients increase evaluation capacity does not justify compromising a needed independent evaluation perspective. For this reason, we see empowerment evaluations as studies that falsely represent un-vetted self-evaluations as defensible evaluations; thus we have labeled such studies as pseudoevaluations. Many commercial entities often provide consumers with a valuable service by acquiring and publishing actual customers' ratings and narrative reviews of products and services purchased. However, sets of such customer feedback may or may not represent complete reporting of an adequate range of candid customer assessments. We see value in customer feedback, especially narrative assessments that allow potential purchasers to see and weigh the criteria that are important to those who took the time to write reviews. Nevertheless, we know from experience that consumers should be circumspect in judging and using such ratings and reviews. We think they should access and use such reviews, but also compare the vendors' reported assessments with those in systematic and defensible consumer product evaluations if such are available. We look more closely at each of these pseudoevaluation approaches in the remainder of this chapter.

The six types of pseudoevaluation we have identified most often occur discretely—that is, each typifies a particular evaluation approach. Unfortunately, pseudoevaluations may combine elements of two or more of these approaches. If this occurs, such procedures become increasingly distant from true, defensible evaluations.

Approach 1: Public Relations Studies

The public relations approach begins with an intention to use data to convince constituents that a program is sound and effective. Other labels for the approach are *ideological marketing, advertising,* and *infomercial*. A public relations study may meet the standard of addressing all right-to-know audiences, but it fails as a legitimate evaluation approach because typically it presents a program's strengths, or an exaggerated view of them, but not its weaknesses.

Clancy and Horner (1999) gave poignant examples of public relations studies that were supposedly but not actually conducted to gain valuable lessons from the 1991 Gulf War:

> In the United States, the Joint Chiefs of Staff and each of the service departments published "Lessons Learned" documents that were in fact advertisements for individual programs, requirements, or services . . . .The so-called "studies" tended to be self-supporting rather than critical of the agency that sponsored the work. And too many of the books, monographs, studies, and official documents misstated the facts, with the aim of salvaging a weapon system, military doctrine, or reputation whose worth could not otherwise be supported. They were public relations documents, not clear-eyed honest appraisals, and they were aimed at influencing the soon-to-come budget reductions and debates over each service's roles and missions. (p. 501)

The advance organizer of the public relations study is the propagandist's information needs. The study's purpose is to help a program's leaders or public relations personnel project a convincing, positive public image of the program. The guiding questions are derived from the public relations specialists' and administrators' conceptions of which questions constituents would find most interesting (and convincing). In general, an evaluator conducting a public relations study seeks information that would most help an organization confirm its claims of excellence and secure public support for a given program. From the start, the evaluator seeks not a valid assessment of the program's merit and worth, but information to help the organization put its best foot forward. He or she avoids gathering or releasing negative findings.

Typical methods used in public relations studies are surveys using biased samples; "push polls" that press respondents to support leading questions designed to garner support for a particular point of view; use of inappropriate norms; biased selection of testimonials and anecdotes; massaging of obtained information; selective release of only positive findings; reporting the central tendency (that is, the average) but not the variation or shape of a distribution of responses; covering up embarrassing incidents; and use of so-called expert advocate consultants. In contrast to seeking honest assessments from "critical friends, public relations studies seek positive endorsements from "friendly critics." A pervasive characteristic of a public relations evaluator's use of dubious methods is a biased attempt to nurture a good picture of a program. The fatal flaw of built-in bias toward reporting only good things offsets any virtues of this approach. If an organization substitutes biased reporting of only positive findings for balanced evaluations of strengths and weaknesses, it soon will demoralize evaluators who are trying to conduct and report valid evaluations and may discredit their overall practice of evaluation.

By disseminating only positive information on a program's performance while withholding information on shortcomings and problems, the evaluator and client may mislead taxpayers, constituents, and other stakeholders concerning a program's true value and what issues need to be addressed to improve it. The possibility of such positive bias in public relations evaluations underlies the long-standing policy of Consumers Union not to include advertising by owners of the products and services being evaluated in its *Consumer Reports* magazine. To maintain credibility with consumers, Consumers Union has, for the most part, maintained an independent perspective and a commitment to identifying and reporting both strengths and weaknesses in the items evaluated and to not supplementing this information with biased ads. (An exception is that the magazine advertises its own supplementary publications and services without presenting clear, independent evaluations of them.)

Evaluators need to be cautious about how they relate to the public relations activities of their sponsors, clients, and supervisors. Certainly, public relations documents will reference information from sound evaluations. Evaluators should do what they can to persuade their clients and other audiences to make honest use of evaluation findings. In this book, we emphasize the importance of clear agreement between evaluator and client on a range of issues, including careful rendering of recommendations. Evaluators should not be party to the misuse of findings, especially when erroneous reports are issued that predictably will mislead readers into believing that a flawed program is effective. As one safeguard, evaluators can promote the use of, and help their clients arrange to engage, independent metaevaluators who will examine the organization's acquisition and use of evaluation findings against professional standards for evaluations (also see Stufflebeam, 1978, 2001c).

Approach 2: Politically Controlled Studies

Politically controlled studies constitute an approach that can be defensible or indefensible. A politically controlled study is illicit if the evaluator or client (1) withholds the full set of evaluation findings from audiences that have an express, legitimate, and legal right to see findings; (2) abrogates a prior agreement to fully disclose evaluation findings; or (3) biases an evaluation message by releasing only part of the findings. It is not legitimate for a client to agree to make the findings of a commissioned evaluation publicly available and then, having previewed the results, to release none or only part of the findings. If and when a client or evaluator violates a formal written agreement on disseminating findings or applicable law, the other party has a right to take appropriate actions or seek an administrative or legal remedy.

An example of a flawed, politically controlled evaluation occurred when a university's president and provost engaged the institution's faculty to evaluate all of its graduate programs. The objective was to identify programs that should be discontinued for such reasons as low demand, low quality, or poor graduation rates. The study's larger purpose was to help the university cut costs in a period of severe fiscal constraints.

The provost reached an agreement with the faculty on a plan for the evaluation, including a guarantee that the full report would be released. The faculty cooperated fully in presenting the needed information and assessing each program against the agreed-on evaluative criteria.

The provost subsequently collected and reviewed the evidence and released her report, which contained her decisions on programs to discontinue but not the evidentiary basis for the cuts.

The faculty protested that the provost's decisions reflected her biases but not the evidence that they had painstakingly collected. When the demanded information was not released, the faculty voted to censure the provost. She resigned the next day, and soon thereafter the university's board of trustees fired the president.

This debacle illustrates the severe costs of corrupt, politically controlled evaluations. Professionals lost their jobs, the incident stimulated much concern in the surrounding community, and many faculty members probably became predisposed to mistrust future proposals for evaluations of their work.

A client sometimes can legitimately commission a covert study and keep the findings private while meeting relevant laws and adhering to an appropriate advance agreement with the evaluator. In the United States, this can be the case for private organizations not governed by public disclosure laws. Furthermore, an evaluator, under a legal contractual agreement, can plan, conduct, and report on an evaluation for private purposes while not disclosing the findings to any outside party. The key to keeping politically controlled studies in legitimate territory is to reach appropriate, legally defensible, advance written agreements and adhere to the contractual provisions concerning release of the studies' findings. Such studies also have to conform to applicable laws on release of information.

The advance organizers for a politically controlled study include implicit or explicit threats that certain political interest groups pose to a program plus the client's need to obtain sensitive information about the program's strengths and weaknesses. The client's purpose in commissioning such a study is to secure assistance in acquiring information of use in improving and defending a program and in combating potential attacks on the program. The questions addressed are those of interest to the client and special groups that share the client's interests and aims. Two main questions are of interest to the client: What is the truth, as best can be determined, surrounding a particular dispute about the program's merits or political attacks on it? What information would be advantageous in a potential conflict situation? Typical methods of conducting the politically controlled study include covert investigations, a focus on selected issues, simulation studies, private polls, reviews of private information records, and (a potential downfall) selective release of findings.

Generally the client of the politically controlled study wants information that is as technically sound as possible. He or she may also, however, want to withhold findings that do not support his or her position, which would push the covert investigation into pseudoevaluation territory. The strength of the approach is that it stresses the need for accurate information. However, because the client might release information selectively to create or sustain an erroneous picture of a program's merit and worth, might distort or misrepresent the findings, might violate a prior agreement to fully release findings, or might violate a right-to-know law, this type of study can degenerate into a pseudoevaluation.

Inappropriate politically controlled studies undoubtedly contributed to the federal and state sunshine laws in the United States and similar laws in other countries. Under current federal and state freedom of information provisions in the United States, most information

obtained through the use of public funds must be made available, in response to an appropriate request, to interested and potentially affected citizens. Thus, there exist legal deterrents to and remedies for illicit, politically controlled evaluations that use public funds. Freedom of information laws are similar in the United States, the United Kingdom, Australia, and a number of other countries. Increasingly over the past few years, freedom of information reports relating particularly to politically oriented or politically motivated reviews have disclosed grossly distorted evaluation reports. The consequences have been extremely embarrassing for the governments or other issuers of the reports. Such disclosures of flawed studies certainly have damaged public confidence in evaluation, accounting, and auditing practices.

Approach 3: Pandering Evaluations

Unfortunately, some evaluators set aside any commitment to the integrity of their evaluation services by catering to a client's desire for certain predetermined evaluative conclusions, regardless of a program's actual performance and outcomes. By delivering the desired conclusion, evaluators often position themselves in the good graces of the client. This can put evaluators in a favored status to conduct additional evaluations for the client in the future.

An example that illustrates the dynamics in pandering evaluations occurred in the context of a federally funded national educational research center. The funding agency required the center to obtain annual external evaluations, and its evaluators would conduct periodic site visits to the center. In those visits, the evaluators would pay special attention to the findings from the center's contracted external evaluations, but they would also see for themselves what was happening in the funded programs.

Year after year, the federal evaluators were perplexed by the apparent lack of progress by one of the center's programs in spite of the highly favorable reports from the center's external evaluator. The program had developed a psychosocial model of child growth and development and had obtained funds each year to validate the model and then apply it to help schools evaluate and improve their elementary school curricula. The federal evaluators saw no evidence during the site visits or in the external evaluator's reports of any work to validate the model and apply it to curriculum development. Instead, the external evaluator applauded the program's work in conducting training sessions on the model and publishing articles about it. Increasingly, the federal evaluators came to the conclusion that the external evaluator was only documenting the amount and quality of training sessions on the model, and that the applauded journal articles were mainly advertisements for the model, devoid of validated findings.

Each year the external evaluator was rehired and deliberated with the center's director and the program's principal investigator to consider how best to persuade federal officials of the program's value. They agreed that funding would be placed in jeopardy if the evaluator reported on the program's omissions and failures in regard to curriculum development and the model's validation. The external evaluator was skillful in preparing impressive, laudatory reports on what the program was doing but ducked the question of whether the center was following through on its full set of commitments. The center's director and the program's

principal investigator were pleased to fund the same external evaluator year after year. In effect, the evaluator each year had bought an evaluation contract for another year by pandering to the client's desires. Eventually the federal agency got wise to this subterfuge; it canceled funding of this program and stipulated that the center find a different, more professional evaluator.

In a pandering evaluation, the evaluator's advance organizers are the client's preferred evaluative conclusions, often leading to a favorable report. The evaluator's immediate purpose is to conduct the evaluation in such a way as to curry and maintain favor with the client; the longer-range purpose is to win future evaluation contracts.

The evaluator and client reach agreement on the questions to be addressed by the evaluation. Often these questions are dictated in the funding agreement covering the program to be evaluated and thus may emanate from a federal agency or other sponsor. The client is not overly concerned about the nature of the evaluation questions but does want to make sure the "right" answers be given, even if they are not true. The client's aim is to obtain a report of positive evaluative conclusions that will pass muster with the funding agency or perhaps a governing board. If some of the funder's questions cannot be finessed, the client and external evaluator may agree to concentrate on the few that can be answered well and hold the others in abeyance.

To obtain the desired conclusions, the evaluator concentrates on those questions whose answers will place the program in a favorable light. The evaluator then employs methods that on the face of it appear to be sound but actually may be biased in their applications. The methods are often manipulated to produce data that appear to support evaluative conclusions. Possible methods are selected anecdotes; push polls; biased samples; biased use of focus groups; carefully selected testimonials; reporting successful cases to the exclusion of failures; arguing that certain questions from the sponsor should be considered later; or presentation of a positive narrative statement.

Pandering evaluations designed to help evaluators buy future contracts from a client have no redeeming features. They may help clients hoodwink their sponsors into believing a flawed program is actually sound, but this is a serious disservice to program sponsors and constituents and to the professional practice of evaluation. These types of practices frequently occur in evaluations of international aid and development programs, for example, where external consultants report positively on program outcomes—often in contradiction to empirical findings—with the hope that they will continue to receive contracts from sponsoring government and nongovernment organizations (Clements, Chianca, & Sasaki, 2008; Cullen & Coryn, 2011; Cullen, Coryn, & Rugh, 2011).

Approach 4: Evaluation by Pretext

Evaluation by pretext exists when an evaluator earnestly and honestly proceeds to conduct a sound evaluation to serve a stated purpose that, unbeknownst to the evaluator, is deceptive and false. In such a case, the client is guilty of the indiscretion of misleading the evaluator. The evaluator is guilty of proceeding with the evaluation without confirming that the evaluation's stated purpose is the actual purpose.[1] The nature of evaluation by pretext is seen in the following true example.

A research center's director had recently been appointed and wanted a baseline evaluation of the center's programs. He contracted for an independent evaluation of the programs. When the evaluation team arrived for its three-day site visit, the center director informed them that the evaluation's purpose should be to identify the full range of flaws in the programs as a basis for program improvement.

He had reviewed previous evaluations of the center's programs and found them reassuring in regard to the high quality of work in the center. These evaluations had been conducted and reported on rigorously and independently. Previous evaluators mainly had found the center's programs to be sound and had lauded them for their importance, rigor, productiveness, and accountability. The director said that although such positive reports had no doubt been good for staff morale, they had not included detailed direction for program improvement. This year he said the evaluators should set aside any search for program strengths; instead, they should concentrate on identifying and cataloging weaknesses.

He said candid reports along these lines would be invaluable to him and the staff for fine-tuning their already good programs and making them truly outstanding. He asked the evaluators to present their findings of program weaknesses at a full staff meeting during the last afternoon of the site visit.

Unfortunately, the evaluators swallowed the director's request and reasoning—hook, line, and sinker. For three days, they delved into each of the center's programs. They were determined to identify and document the full range of weaknesses in each program.

At the end of their visit, the evaluators went to the auditorium where they would orally deliver the findings to the center's staff. To the surprise and consternation of the evaluators, the audience included not only the center's staff but also officials from the federal agency that was funding the center's work. The evaluators wished they were in a position to present a balanced assessment of each program's strengths and weaknesses because the center's funding undoubtedly was at risk. However, they were prepared only to present what they had searched for and found: program weaknesses. The evaluators' recitations on program weaknesses cast a pall over the entire meeting and undoubtedly misled the federal officials as to the true merit and worth of the programs being reviewed.

Why would the center's director orchestrate such a disastrous chain of evaluation events? It turned out that he had not liked the previous director of the center, wanted to discredit his leadership, and was seeking to replace the center's programs with others of his choice. These were the evaluation's real purposes as viewed by the director. In this evaluation by pretext, the evaluators had unwittingly played into the director's hands. With some advance exploration before signing on to do the evaluation, they might have learned of the director's deception and declined the evaluation assignment or insisted on making a valid assessment of strengths and weaknesses.

Before contracting for an evaluation, it is a good idea to obtain information from a wide range of program stakeholders. It can be particularly enlightening to consider who might be hurt by the evaluation and to invite their reactions. In interacting with stakeholders, prospective evaluators should outline what they have been asked to do and inquire as to what concerns they should consider before agreeing to conduct an evaluation. Also, evaluators probably should not

agree to collect and report only strengths or only weaknesses in a program. As stated in the 1994 and 2011 editions of the Joint Committee on Standards for Educational Evaluation's *Program Evaluation Standards,* evaluators should fairly appraise both strengths and weaknesses in a program.

The main advance organizer in an evaluation by pretext is the client's directive to the evaluator and rationale for the directive—for example, to identify program defects as a basis for program improvement. The client's purpose is not the purpose given to the evaluator. In the example here, the director's purpose was not program improvement, as stated to the evaluators, but program termination and discrediting of the previous director. Clients are the source of evaluation questions that guide evaluations by pretext. Although clients should be one source of evaluation questions, they should not be the only source: other stakeholders and evaluators themselves should also contribute evaluation questions. Typical questions in evaluations by pretext focus on negative aspects of a program, but they may be more varied depending on the evaluation's hidden purpose and could concentrate on only positive features of a program. Because evaluations by pretext employ and are led astray by the evaluation questions stated by the client, methodology is not the source of these evaluations' problems. This approach to evaluation has no redeeming qualities and can be seen as disturbingly Machiavellian.

Approach 5: Empowerment Under the Guise of Evaluation

When an external evaluator's efforts to empower a group to conduct its own evaluations are advanced as external or independent evaluations, they fit our conceptualization of empowerment under the guise of evaluation.[2] Such applications of empowerment evaluation give evaluees and other program stakeholders the power and authority to write or edit interim or final reports while claiming or giving the illusion that an independent evaluator prepared and delivered the reports—or at least that he or she endorsed internal evaluation reports. In such cases, an external evaluator is preoccupied with developing rapport with and assisting a vulnerable or disadvantaged group whose work is to be evaluated. The empowerment evaluator's central objectives are to help a group of evaluees maintain and increase resources, train them in evaluation, empower them to conduct and use evaluation to serve their interests, or lend them sufficient credibility to make conducted evaluations influential. The external empowerment evaluator serves as a critical friend (or, more likely, a friendly critic). The umbrella approach described here largely was developed and advocated by Fetterman (1994, 2001; Fetterman & Wandersman, 2005).

Objectives of training and empowering a disadvantaged group to conduct evaluations are laudable in their own right, and fostering self-determination is the defining focus and heart of empowerment evaluation's explicit political and social change agenda. However, empowering groups to do their own evaluations is not professional, disciplined evaluation (also see Donaldson, Patton, Fetterman, & Scriven, 2010; Patton, 2005a; Scriven, 1997, 2005d; Stufflebeam, 1994, 2001b). It is empowerment by such evaluation capacity development activities as offering evaluation training; developing an organization's evaluation policies, procedures, and tools; or setting up an office of evaluation services. Empowerment activities

move into the pseudoevaluation range when an external evaluator credits an internal evaluation as his or her own, credits a flawed internal evaluation as sound, stands silent when the client attributes the evaluation's findings to the external evaluator, or fails to ensure that the evaluation will be subjected to an independent metaevaluation. An actual example of empowerment under the guise of evaluation follows.

A government organization in an African country engaged an American researcher to evaluate an educational improvement program being funded in a remote, primitive area of the country. The evaluator quickly found that the program's funds were mainly going into graft, that there was no discernible effort to implement the agreed-on educational reforms, and that no objectives had been achieved. At this point, the evaluator concluded that the program was a total failure. He also realized, however, that the program's target area was poverty stricken and that the situation in the area would only worsen if the government stopped funding the program.

The evaluator decided not to write and submit a report exposing the program's failure. Instead, he chose to redefine his task as empowerment evaluation. Accordingly, he offered evaluation training to area personnel and collaborated with them to produce their own evaluation of the government-sponsored program. The resulting evaluation report was highly positive. The external evaluator acquiesced in the group's submission of the favorable evaluation report to the funding agency and even showed his support by writing a preface to the report. In that preface, he stated that the program's staff were to be congratulated for their development of evaluation capacity and their production of an informative evaluation report. In a private conversation with this external evaluator, he related that the good he had done by helping keep government funding in the poverty-stricken area far outweighed his transgression of endorsing a faulty evaluation report and allowing it to go forward.

The claimed short-range benefits of this empowerment evaluation experience were outweighed, however, by the external evaluator's having conveyed very bad lessons to the program's staff: (1) biasing an evaluation report is acceptable practice if it helps secure a desirable end; (2) it is acceptable to ask independent evaluators to serve as advocates and give clients control of the evaluation work; (3) program personnel with little or no evaluation expertise should trust their own assessments of their own work; (4) credibility for biased self-evaluations can be bought by selecting the "right" external evaluation expert; and (5) it is unnecessary to subject internal evaluations to credible, independent metaevaluations. All of these lessons teach clients to continue engaging in corrupt, incompetent evaluation practices.

The advance organizer in studies employing empowerment under the guise of evaluation is an external evaluator's dedication to helping a group of evaluees or other program stakeholders gain power to improve the group's situation. The purposes of this type of pseudoevaluation are to empower group members to conduct their own evaluations and lend credibility to those evaluations. The questions for such evaluations usually are stipulated by external funding organizations. Typical questions concern whether funded programs are being implemented as promised and whether they are succeeding, most likely from the points of view of staffing strength and positive activities. The external evaluator's method is mainly to provide on-the-job training and technical support to help the evaluees conduct their own evaluations. The external evaluator's efforts become corrupt when he or she inappropriately endorses or shares credit for evaluation

reports produced by the evaluees or allows the evaluees to falsely attribute evaluation findings to the external evaluator. This type of evaluation has no strengths because it can aid and abet groups in putting forth faulty evaluations; it reinforces the notion that subterfuge in evaluation is acceptable; it teaches corrupt evaluation practices; it is not subjected to credible, independent metaevaluations; and it contributes to discrediting the professional practice of evaluation.

A strict effort to help groups develop evaluation capacity is, of course, commendable. Steps to reduce evaluees' fear of evaluation (see also Donaldson, Gooler, & Scriven, 2002); train them in evaluation concepts and methods; and involve them in undertaking evaluation work are in the interest of mainstreaming evaluation. But such constructive, capacity-building steps do not themselves constitute evaluation. When evaluation capacity building is labeled "evaluation," as in empowerment evaluation, it is false advertising, because such efforts do not amount to evaluation. Helping staffs develop sound evaluation capacity absolutely requires them to subject their evaluations to credible, independent metaevaluations, a practice that is alien to the precepts of empowerment evaluation. We believe that requiring empowerment evaluations to be independently evaluated would doom the approach's survival or force radical changes in its orientation and application.

An evaluator must not give evaluees power over an external evaluation message, even in the interest of reducing their fear of and antipathy toward evaluation. The often predictable result of empowerment under the guise of evaluation is essentially a biased self-report that masquerades as an unbiased, independent evaluation. Moreover, this is modeling of bad evaluation work; accordingly, evaluees are empowered not to conduct rigorous, creditable evaluations, but to make a game of what should be a sound evaluation enterprise.

In 2006 Miller and Campbell applied Wandersman et al.'s ten principles of empowerment evaluation (2005; see Table 5.1) to analyze and assess the processes and outcomes of forty-seven published examples of empowerment evaluation. The results of their review suggested that empowerment evaluators' adherence to the ten principles in practice widely varied; that there was a lack of credible evidence demonstrating empowered outcomes; and that some of the criticisms (for example, Patton, 2005a; Scriven, 1997; Sechrest, 1997; Stufflebeam, 1994, 2001b) leveled against the approach were warranted.

Approach 6: Customer Feedback Evaluation

For the customer feedback evaluation approach, we have in mind the common practice by numerous commercial enterprises of obtaining and publishing on their Web site consumers' ratings and (sometimes) narrative reviews of products or services they had supposedly purchased and used. In some fields this approach is referred to as "word of mouth" (Dellarocas, 2003; Hennig-Thurau, Gwinner, Walsh, & Gremler, 2004; Resnick, Zeckhauser, Friedman, & Kuwabara, 2000). We have mixed feelings about treating this approach as a form of pseudoevaluation, because we, like many consumers, appreciate the opportunity to see what purchasers had to say about a product or service that we are considering buying. That being said, users of reported customer feedback usually have to take it on faith that the feedback is sound, and often it isn't. Witness, for example, multiple "customer comments" that are the same, word

Table 5.1 Principles of Empowerment Evaluation

Dimension	Principles
Process	• A community should make the decisions about all aspects of an evaluation, including its purpose and design; a community should decide how the results are used (community-ownership principle). • Stakeholders, including staff members, community members, funding institutions, and program participants, should directly participate in decisions about an evaluation (inclusion principle). • Empowerment evaluations should value processes that emphasize deliberation and authentic collaboration among stakeholders; the empowerment evaluation process should be readily transparent (democratic-participation principle). • The tools developed for an empowerment evaluation should reflect community wisdom (community-knowledge principle). • Empowerment evaluations must appreciate the value of scientific evidence (evidence-based-strategies principle). • Empowerment evaluations should be conducted in ways that hold evaluators accountable to programs' administrators and to the public (accountability principle).
Outcomes	• Empowerment evaluations must value improvement; evaluations should be tools to achieve improvement (improvement principle). • Empowerment evaluations should change organizations' cultures and influence individual thinking (organizational-learning principle). • Empowerment evaluations should facilitate the attainment of fair allocations of resources, opportunities, and bargaining power; evaluations should contribute to the amelioration of social inequalities (social-justice principle). • Empowerment evaluations should facilitate organizations' use of data to learn and their ability to sustain their evaluation efforts (capacity-building principle).

Source: Adapted from Miller, R. L., & Campbell, R. (2006). Taking stock of empowerment evaluation: An empirical review. *American Journal of Evaluation, 27,* 300.

for word. There is no aspect of sampling theory to substantiate such an occurrence. A more likely hypothesis—or, at least, suspicion—is that the vendor manufactured the responses.

The advance organizers of the customer feedback approach typically are consumer-provided ratings of a product or service, often based on a five-point rating scale; respondents' comments; and possibly the vendor's responses to questions submitted by customers. Ostensibly, the main purpose of the customer feedback approach is to obtain and report candid feedback from purchasers of a product or service. However, depending on whether all obtained feedback is released or comments are indeed from actual consumers, the purpose could be to project a positive image of the product or service. The source of questions addressed by customers is the vendor, not the target audience of consumers. Questions asked of consumers may include, How do you rate the quality of the product or service on a five-point scale? What are the pros of the product or service? What is the best use of the product or service? What type of consumer are you? and What is your bottom-line recommendation in regard to the product or service? The typical method used is inviting purchasers to open and complete a Web-based questionnaire. The reported results for a given product or service may include the number of reviews; the average rating; specific pros identified and the number of reviewers who cited each one; and verbatim comments by each reviewer who offered comments, including the reviewer's bottom-line judgment or recommendation. We know of no particular pioneers for this approach nor of any systematic investigations of its practice and consequences, and it

is, therefore, ripe for empirical study. Even so, its application has been widespread since the advent of Internet marketing and selling of consumer products and services.

This approach does, however, have a potential strength when a vendor obtains and reports comments as well as ratings. On the one hand, if they come from actual consumers, comments can be invaluable in helping customers consider whether the customer feedback reflects criteria that are important to them in deciding to purchase or not purchase a particular item. On the other hand, if the comments are manufactured and not from actual consumers, potential customers can be seriously misled as to the quality of the product or service.

Despite the approach's clear attractiveness and utility, three problems with it led us to include it in the pseudoevaluation category. The first is that some customer feedback—for example, that reported by certain movie rental services—only includes the average number of stars (⋆) on a five-star basis that respondents assigned to a particular movie (and sometimes the number of ratings). Although such information is of interest, it begs the question of what criteria purchasers used to assign their ratings. And it embodies our second reservation about this approach, which is that little information is available about the characteristics and representativeness of the persons who provided ratings, or whether they actually purchased or used the product or service. Information might or might not have been provided by a reasonably broad sample of supposed purchasers, but more likely not, and one cannot make a judgment on this. As noted earlier, customer feedback reports are enhanced when they include both ratings and narrative comments; the comments can be especially useful because typically they reveal the criteria that each respondent used to judge a product or service. Even then, however, the usefulness of this type of feedback is limited because of the third problem we see with this approach: one does not know whether the vendor reported all obtained feedback or whether all the reported comments came from actual consumers who chose to provide legitimate feedback. Although we support the practice by commercial enterprises of reporting customer feedback, we urge consumers to contrast the feedback with more credible assessments of products and services, if such can be found. Moreover, we think vendors' practices of collecting and reporting customer evaluations should from time to time be validated via sound procedures of metaevaluation.

Overall, we believe that vendors should continue to obtain and report ratings and purchasers' valid comments on products and services, along with information about raters. Moreover, it would be appropriate for vendors to certify on their Web site that the reported information is a complete account of obtained customer feedback. Vendors also should consider periodically obtaining and reporting the findings of an independent assessment of their use of the customer feedback approach. Potential purchasers are advised to compare reported customer feedback with the more systematic evaluations of products or services that often may be available from the *Consumer Reports* Web site (www.consumerreports.org) or other independent, credible Web sites providing such evaluations.

Summary

Although it would be unrealistic to recommend that administrators and other evaluation users not obtain and selectively employ information for maintaining political or economic viability, evaluators should not lend their name and endorsement to evaluations presented by their clients

that misrepresent a complete set of relevant findings, contain falsified information aimed at winning political contests, or violate applicable laws or prior formal agreements concerning the release of findings. If evaluators acquiesce to and support such pseudoevaluations, they help promote and support injustice, mislead decision making, project an erroneous concept of evaluation, lower confidence in evaluation services, and discredit the evaluation profession. Even when an evaluator's objective is socially constructive, nothing worthwhile is achieved by empowering groups to conduct their own evaluations if they are essentially taught that biased self-reports, erroneously credited as independent evaluations, are acceptable. We do note that evaluators can give private evaluative feedback to clients legitimately, provided that the evaluation is sound and conforms to pertinent laws, statutes, and policies as well as appropriate contractual agreements on the editing and release of findings. Also, we encourage Internet-based vendors to continue obtaining and reporting customer feedback, but to assess and make transparent the extent to which such feedback comports with standards of the evaluation field.

REVIEW QUESTIONS

1. How do you define pseudoevaluation?
2. Give two examples of politically controlled studies.
3. Name the six pseudoevaluation approaches identified and discussed in this chapter, and give a one-line explanation of each.
4. List the main flaws of (a) politically controlled studies, (b) pandering evaluations, (c) empowerment under the guise of evaluation, and (d) customer feedback evaluation.
5. A distorted, overly positive view of a program is released to the public, while problematic facts are withheld. Which form of pseudoevaluation applies to this statement, and why?
6. An evaluation client is actually unconcerned about an evaluation's stated questions and obtained evidence, but surreptitiously plans to use the existence of the evaluation to justify taking certain unannounced but already planned actions. Which form of pseudoevaluation applies to this characterization, and why?
7. In the hope of obtaining future evaluation assignments, an evaluator agrees to investigate a program, but to report only positive findings. What form of pseudoevaluation is this, and why should the "evaluator" be discredited?
8. A company director engages an independent evaluator to prepare her employees to evaluate one of the company's programs. The evaluator provides the employees with a brief training session on sound evaluation procedures. The employees then evaluate the selected program; produce a final, highly positive evaluation report; and identify the external evaluator as the report's principal author. With the external evaluator's concurrence, the company director then sends the report to the program's outside funding agency. What form of pseudoevaluation is this, and what are its main flaws?

9. A public school district's superintendent directs the district's evaluator to maintain accurate information on the strengths and weaknesses of each school in her district, but to report that information only to the superintendent. The superintendent also authorizes the evaluator to include only the positive or "nonsensitive" information in reports to the public. What form of pseudoevaluation is this, what are its strengths and weaknesses, and in what ways is the superintendent acting inappropriately and probably illegally?
10. What are the pros and cons of the customer feedback approach to evaluation, and what provisions are needed to legitimize the use of this approach?

Group Exercises

Work through the following two exercises with your group. It is quite possible that members will reach different conclusions about how best to respond to the problems. However, members should try to justify their position.

Exercise 1

Discuss the rationale that supports the following statement: "Evaluators should not lend their name and endorsement to evaluations presented by their clients that misrepresent the full set of relevant findings."

Exercise 2

Truth is often stranger (and perhaps more disconcerting) than fiction. The following situation is based on fact.

A large mining company in the northwestern part of Western Australia depended heavily on both federal government subsidies to ensure strong exports as well as an annual report that was received favorably by shareholders. The same firm of evaluators (incorporating auditors) examined the company from 1996 to 2003. Reports to the federal government and shareholders over this time span were produced in a glossy format; invariably, the organization was portrayed as flourishing. As a result of questions being asked at the 2003 annual general meeting of shareholders (with government representatives present), a subsequent independent evaluation found an abundance of corruption, significantly hiding massive financial losses for the previous six years.

The deceptions of the original firm of evaluators were characterized by

- Tight control of the kinds of information released, influenced strongly by both the senior administrators of the company being evaluated as well as federal government officials (who were determined to pursue a favorable balance of trade in mining products with Asian countries)

- A consistent desire to give a glowing annual report, fully knowing that there was illegal collusion with mining management and government officials about the questions to be addressed annually and which matters would be omitted
- Data and information manipulated to exaggerate preconceived positive outcomes

In this sorry saga, who was at fault, and why? What advice would you give to the mining company and the federal government for future evaluations?

Notes

1. Patton (2008) described numerous methods for identifying and avoiding such perils using situational analysis and other techniques.
2. Empowerment evaluation shares an emphasis on social justice with the social agenda and advocacy approaches described in Chapter 8. However, we classified empowerment evaluation not as a social agenda or advocacy approach, but as a pseudoevaluation approach, because it cedes authority over key matters relating to controlling an evaluation's quality to the subject program's stakeholders, and because a wide range of reviewers of the approach have judged it to be grossly deficient in relation to the standards of the evaluation field.

Suggested Supplemental Readings

Dellarocas, C. (2003). The digitization of word of mouth: Promises and challenges of online feedback mechanisms. *Management Science*, *49*, 1407–1424.

Fetterman, D. M. (1994). Empowerment evaluation. *Evaluation Practice*, *15*, 1–15.

Fetterman, D. M. (2001). *Foundations of empowerment evaluation: Step by step*. Thousand Oaks, CA: Sage.

Fetterman, D. M., & Wandersman, A. (Eds.). (2005). *Empowerment evaluation principles in practice*. New York, NY: Guilford Press.

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, *18*(1), 38–52.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Joint Committee on Standards for Educational Evaluation. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Miller, R. L., & Campbell, R. (2006). Taking stock of empowerment evaluation: An empirical review. *American Journal of Evaluation*, *27*, 296–319.

Resnick, P., Zeckhauser, R., Friedman, E., & Kuwabara, K. (2000). Reputation systems. *Communications of the ACM*, *43*(12), 45–48.

Scriven, M. (1997). Empowerment evaluation examined. *Evaluation Practice*, *18*, 165–175.

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.

CHAPTER 6

QUASI-EVALUATION STUDIES

Quasi-Evaluation Approaches Defined

A quasi-evaluation approach provides direction for performing a high-quality study that is narrow in terms of the scope of questions addressed, the methods employed, or both. Such a study may be narrow because it (1) focuses on one or more limited questions, such as Were the developer's objectives achieved? (2) employs only one selected method, such as administration of a questionnaire to a sample of program participants; or (3) employs only one method and also addresses only one or a few narrowly focused questions (which is often the case with experimental studies).

The questions that drive a quasi-evaluation might be derived, for example, from a program's stated objectives; a newspaper's decision to report rankings of schools on a single variable of interest (such as average reading test scores); a funding agency's requirement that contractors submit periodic reports of accomplishments; or a focused investigation to find, document, and publicize either the strengths of an apparently successful program or the weaknesses of a program perceived to be failing. A quasi-evaluation study that starts with a selected method might employ as its starting point a design for a randomized controlled experiment, a particular standardized test, a cost analysis procedure, or a program theory. Such approaches tend to emphasize technical quality, as illustrated by studies based on an experimental design or a standardized test and set of norm tables. In general, proponents of quasi-evaluation approaches stress that it is usually better to answer a few pointed questions promptly and well (that is, validly

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **The definition of a quasi-evaluation approach**
- **The functions of quasi-evaluation approaches**
- **The strengths and weaknesses of quasi-evaluation approaches**
- **The identity and characteristics of eight quasi-evaluation approaches: objectives-based studies, the Success Case Method, value-added assessment of outcomes, experimental and quasi-experimental studies, cost studies; connoisseurship and criticism, theory-based evaluation; and meta-analysis**[1]

and reliably) than to attempt a longitudinal assessment with sufficiently broad scope to comprehensively and convincingly assess a program's merit and worth (also see Cronbach & Snow, 1981).

Functions of Quasi-Evaluation Approaches

In an ideal world, every evaluation would provide information of sufficient scope and quality to fully assess a program's merit and worth. Although this is a worthy goal for many program evaluation assignments, it is often unrealistic or excessive considering the ebb and flow of clients' specific, immediate evaluative information needs. Clients might need only a focused, factual response to a specific question. They might want and need only

- A needs assessment report to support a scheduled budget hearing
- Confirmation that a funded contractor is meeting basic contract requirements and doing so on time
- A summary of current findings related to each objective of a funded program
- A study comparing experimental and control groups to determine if a treatment's effect on an outcome variable was statistically superior to the results for the control group
- A quick troubleshooting study to determine why a program apparently is in chaos

All of these examples reflect clients' real-world requirements for evaluative information. All of them fall short of calling for a study to fully assess a program's merit and worth. But evaluator responses to such narrow requests for evaluative information are not only frequent but also important and valued by clients. When conducted well, quasi-evaluation studies deliver a valuable service in providing high-quality, timely responses to requests for targeted, limited evaluative feedback.

General Strengths and Weaknesses of Quasi-Evaluation Approaches

We see quasi-evaluation studies as legitimate, necessary, and useful in their own right. Their most important strengths are (1) an orientation toward responding directly, in a timely fashion, and at a high level of quality to client needs; (2) efficiency in collecting only the information that the client requests or that the selected procedure requires; and (3) a tendency to yield information that has high levels of validity, reliability, credibility, and immediate utility.

The main weakness of quasi-evaluation studies is in the unchecked possibility for clients mistakenly to believe or represent that such studies constitute thorough evaluations of a program's merit and worth. In most cases a quasi-evaluation study's focus is too limited to address the full range of questions pertaining to merit and worth. Such an approach is unlikely to meet all the requirements of a sound evaluation, as defined in professional standards for program evaluations. We see that shortcoming as a problem and weakness of the

quasi-evaluation study if a program's leader or other interested party misrepresents such a study as being a comprehensive evaluation of a program's merit and worth. We present this caveat because we have encountered a few such cases of misrepresentation. We advise evaluators who conduct quasi-evaluations, and their clients, to make clear the limitations of judgments that may legitimately be reached based on the obtained findings. Of course, this advice applies to all types of evaluations.

Approach 7: Objectives-Based Studies

The objectives-based study is the classic example of a quasi-evaluation approach that focuses on a narrow set of questions. Madaus and Stufflebeam (1988) provided a comprehensive look at this approach in an edited volume of the classical writings of Ralph W. Tyler.

Advance Organizers

In this approach, some statement of a program's objectives constitutes the advance organizer. The defined objectives provide the basis for determining what information should be collected. The specified information needs provide direction for identifying pertinent information sources and developing or selecting tools to measure program recipients' performance in relation to each objective. When program objectives are defined in great detail, they include specification of cut scores above which a program is judged to have met its objectives.

Purposes

The usual purposes of an objectives-based study are to specify and define clearly what a program is intended to accomplish and to determine the extent to which the program achieved its objectives. Program administrators use the results of objectives-based studies to report on the extent to which their program delivered the promised outcomes. Objectives-based program reports also allow clients and program recipients to reach their own evidence-based judgments of a program's level of success. Program staffs may use judgments of which objectives were not achieved as diagnostic feedback for use in rethinking and improving a program's design and execution. Funders may use results of objectives-based studies to sustain funding, increase or decrease funding, or terminate a program.

Sources of Questions

This approach's general question is, Did a program achieve its objectives? The objectives may be defined by a program's staff, defined and mandated by the program's funder, or formulated and explicated by the evaluator in consultation with the program's staff. The validity of the objectives resides mainly in their acceptability to the program's staff, beneficiaries, and funder.

Questions

The specific question addressed by objectives-based studies is, To what extent did the program achieve each of its stated objectives? The program's objectives are expected to be defined

in very clear terms. In some studies, each program objective is defined in such detail that it specifies what is to be achieved, the conditions under which the achievement is to be produced, how the achievement is to be measured, and the level on the pertinent measurement scale that is to be counted as successful.

Methods

Typically, but not always, an objectives-based evaluation is an internal study done by a developer or other program leader or, less often, by a program service provider. The methods used in objectives-based studies essentially involve specifying operational objectives and collecting and analyzing pertinent information to determine how well each objective was achieved. R. W. Tyler (for example, 1932, 1942, 1950, 1966, 1967) stressed that a wide range of objective and performance assessment procedures usually should be employed. This sets his approach apart from studies that focus on a particular method, such as an experimental design or a single standardized test. Criterion-referenced tests and students' work samples are especially relevant to the objectives-based approach.

Pioneers

Tyler is generally acknowledged to be the pioneer of the objectives-based type of study, although Percy Bridgman and Edward Thorndike should also be credited. Several people have developed variations of Tyler's model. They include Bloom, Englehart, Furst, Hill, and Krathwohl (1956); Hammond (1972); Metfessel and Michael (1967); Popham (1969); Provus (1971); and Steinmetz (1983). Although Tyler developed the objectives-based approach for use in evaluating educational programs, this approach's influence has spread far beyond the confines of education. Objectives-based evaluations can be found in virtually all fields of service, and it is common to see government requirements specifying that evaluations be conducted to determine the extent to which each funded program achieved its objectives.

Use Considerations

The objectives-based approach is especially applicable in assessing tightly focused programs that have clear, supportable objectives. Even then, such studies can be strengthened by judging program objectives against intended recipients' assessed needs, searching for side effects, and studying the process as well as outcomes. In practice, it is rare for evaluators to question program objectives (Scriven, 1974, 1991).

Strengths

Objectives-based investigation has been the most prevalent approach to evaluating programs. Perhaps this is due to the approach's ease of application. It has commonsense appeal; program administrators have had a great deal of experience with it; and it makes use of published rules for

writing operational or behavioral objectives, both norm-referenced and criterion-referenced testing, and performance assessments.

Weaknesses

Common criticisms are that objectives-based studies report findings only at the end of a program; that such information is neither timely nor pertinent to improving a program's implementation; that the information often is far too narrow to constitute a sufficient basis for judging a program's level of success, especially for the full range of beneficiaries; that objectives-based studies do not uncover positive and negative side effects; that they may credit unworthy objectives; and that they fall short of assessing a program's significance (see Scriven, 1974).

Approach 8: The Success Case Method

A recent entry in the lexicon of quasi-evaluation approaches is the Success Case Method, developed by Robert Brinkerhoff (2003, 2005a, 2005b, 2006). In this approach, the evaluator deliberately searches for and illuminates instances of success and contrasts these with what is not working in a program.

Advance Organizers

The advance organizers in this approach are a program's specific successes; how they were produced; how the most impactful instances of success compare to the least successful instances (for example, in regard to prevalence and importance); and the contextual circumstances and other causal factors that contributed to the successes. These are aspects of a program that, when identified and substantiated, would be important in sustaining, expanding, or improving the program.

Purposes

The purpose of this approach is to provide change leaders with a simple, dependable, and low-cost way of expeditiously finding out how well and in what respects a change effort is working. The intent of the Success Case Method is to discover, analyze, and document any successes a program might be having so they can be built on and extended (assuming that these successes are worthwhile). The Success Case Method is put forward not as a comprehensive approach to fully assessing an enterprise's merit and worth over time (however, see Coryn, Schröter, and Hanssen [2009] for an example of a longitudinal Success Case Method), but as a relatively quick yet defensible means of gathering critically important information for use in program improvement. The approach may be employed in conclusion-oriented summative evaluations, but mainly it is intended for use in formative evaluations aimed at program improvement.

In responding via an e-mail message to a previous draft of this characterization of the Success Case Method for this book's first edition (Stufflebeam & Shinkfield, 2007), Brinkerhoff stated:

> No program is ever wholly successful or unsuccessful; thus methods that look for "average" or typical outcomes inevitably underestimate strengths and overestimate weaknesses. The Success Case Method helps evaluators capture the successes and then assess their worth. If the best that a program is doing is not good enough, the Success Case Method evaluation is finished (and so, usually, is the program). But if the good stuff is indeed worthy, then it may make sense to get a greater return on the program investment by leveraging the strengths, which means as well that we have to figure out WHY it works WHEN it works; this is also an aim of the Success Case Method (and is why we always compare the least impactful instances to the most impactful ones).

Sources of Questions

Questions about where to look for successes (and failures) often are identified by the people who are most directly involved in carrying out a program or receiving its services. When traditional methods of evaluation have branded a program as unsuccessful, persons closely associated with the program may believe there are valid reasons to dispute the conclusion. Accordingly, they may put forward their perspectives, cite their associations with program successes, and present hypotheses of program strengths that could and should be confirmed through further investigation.

Questions

General questions addressed by the Success Case Method include the following:

- What are the noteworthy successes of the given program?
- How were the program's successes produced?
- What contextual and other causal factors contributed to the program's successes?
- How important are the identified successes as bases for further program development?
- How do the program's most impactful features compare with its least impactful features?

Methods

The Success Case Method, as described by Brinkerhoff (2003), typically is conducted using a five-step procedure:

1. Focus and plan the Success Case Method.
2. Create an impact model.
3. Survey all program recipients to identify success and nonsuccess cases.

4. Interview key informants in a random sample of success and nonsuccess cases and document their stories.
5. Communicate findings, conclusions, and recommendations.

The focusing and planning of a Success Case Method study (in step 1) can take many forms. As already mentioned, the approach can be used for both formative and summative evaluation purposes, although the approach is typically used for formative purposes. Once the focus of the study has been determined, (in step 2) an impact model is developed that delineates how an intervention is assumed to produce its desired results (see "Approach 13: Theory-Based Evaluation" in this chapter for a discussion relevant to applying program theories to study program processes and impacts). Then (in step 3) cases are identified as high (H), or success cases; moderate (M), or average cases; or low (L), or failure or nonsuccess cases, or some similar variation. Typically cases are classified using survey methods specifically designed to provide information for classifying cases (that is, some measure of success, such as return on investment; also see "Approach 11: Cost Studies"). Once classified, (in step 4) these cases serve as sampling strata, and cases are randomly selected from the upper and lower ends of the success measure. The Success Case Method is, therefore, essentially an analysis of extreme or outlier cases, as opposed to average cases, whereby independent evidence is sought to corroborate claims of success or failure (for example, to determine whether a salesperson's sales actually increased following an in-service training experience). Also in step 4, the reasons underlying successes or failures are investigated using semistructured interview techniques designed to probe possible explanations from a random sample of extreme cases. Finally, (in step 5) findings, conclusions, and recommendations are communicated. Often, evaluation reports using the Success Case Method are presented in the form of "success stories" (Coryn et al., 2009, p. 81). Figure 6.1 illustrates the basic logic in applying the Success Case Method where an observable effect has occurred and cases have been classified as high, medium, or low.

In Success Case Method evaluations, it is assumed that the sample of successful and unsuccessful cases is normally distributed over the success measure as shown in Figure 6.2. Also assumed is the cases at the extremes—the focus of the Success Case Method—of such a distribution (for example, between $+2\sigma$ and $+4\sigma$ or $\geq +3\sigma$ [that is, success cases] and between -4σ and -2σ or $\leq -3\sigma$ [that is, failure cases]) also are normally distributed. Such assumptions, however, are rarely met in most program evaluation situations (for example, there might be nonsymmetric distributions of success).

Pioneers

Brinkerhoff (2003) has been the pioneer in conceptualizing, applying, and publicizing the approach. Scholars whom he credits with influencing his development of the Success Case Method include Egon Guba, Barry Kibel, Annette Simmons, and Robert Stake.

Use Considerations

Although the Success Case Method has been applied primarily in the for-profit sector to evaluate training initiatives or new work methods (Brinkerhoff, 2005), it recently has been

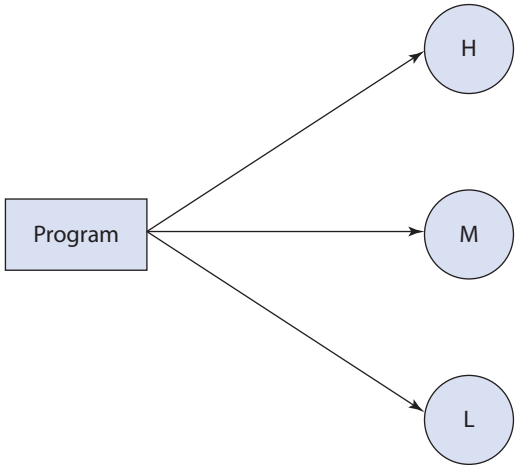

Figure 6.1 Conceptual Model of the Success Case Method

Source: Coryn, C.L.S., Schröter, D. C., & Hanssen, C. E. (2009). Adding a time-series design element to the Success Case Method to improve methodological rigor: An application for nonprofit program evaluation. *American Journal of Evaluation, 30,* 81.

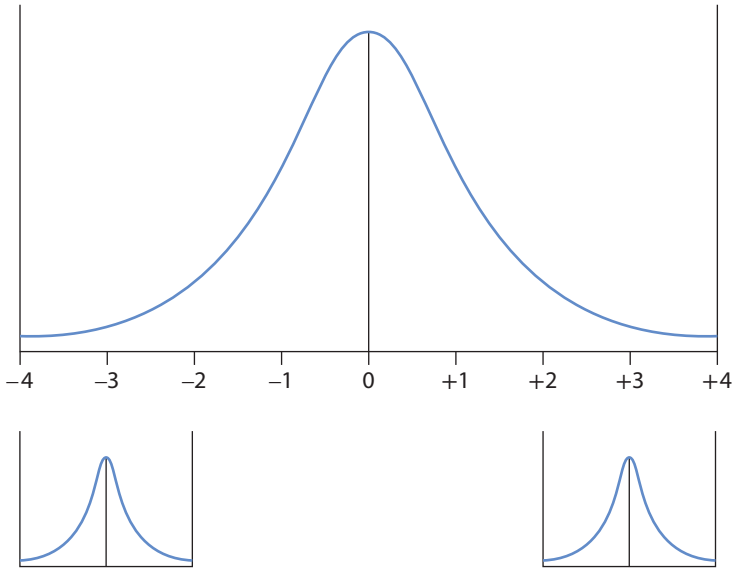

Figure 6.2 Distributional Assumptions of Success Case Method Samples

applied to an evaluation of the impact of a U.S.-based and international food security project (Chianca & Risley, 2005); in educational settings to determine the factors that influence the academic achievement of minority students (Coryn, Schröter, et al., 2007); to evaluate organizational learning in a nonprofit foundation (Berkley, Day, Smith, & Chianca, 2005); and to evaluate an assistance initiative for persons who are homeless and those at risk of

homelessness, an assessment that was taken up within the effectiveness evaluation component of a context, input, process, and product (CIPP) model evaluation conducted by Coryn et al. (2009; also see Chapter 13).

We think the following example illustrates the intent, spirit, and procedure of the Success Case Method.

An evaluator had been contracted to evaluate a highly funded vocational education program that was designed for widespread implementation in secondary schools in Ohio. In accordance with the evaluation contract, the evaluator had conducted a comparative experiment. Schools across Ohio had been randomly assigned to receive the new program or not. Following implementation of the program, the evaluator had compared the two groups on tests of knowledge of vocational education content and other measures of attitude and aspiration. Across all of the outcome measures, the consistent finding was that there were no statistically significant differences between students in the group that had received the program and students in the control schools. The expensive program was thus judged a failure and not worthy of continuation and widespread implementation throughout the schools of Ohio.

Release of these findings brought protests from the schools in the experimental group. Teachers in those schools said the program had made substantial and important impacts on their students, even if the evaluators had been unable to detect these impacts. The teachers said they had seen the impacts with their own eyes, and they worried that this less-than-sensitive evaluation would lead to the termination of a meritorious program that had good value for students throughout the state.

These teachers were so persuasive that the evaluator decided to take another look at the data (through what we might today term a "success case study"). He wanted to ascertain whether he could find convincing evidence that the teachers were correct about the program's having made important impacts on their students. Using an item analysis procedure, the evaluator searched for test items that discriminated statistically between experimental and control students. He was surprised at the results of this search. He found a sizable number of items on which students in the experimental schools outperformed the students in the control schools. He found another, small subset of items on which the control students actually outperformed the students in the experimental group. And he found a third set of items that did not differentiate between the two groups. Content analyses of the test items on which the experimental students excelled confirmed what the teachers had been reporting as successes in their classrooms. Moreover, the items that showed superior performance for the experimental group were judged to reflect important impacts. The few items that showed superior responses from students in the control group were deemed important for further investigation, as were the items that showed no differences between the two groups.

The original comparative analysis that combined all of the items in each test had obscured the important underlying statistical interactions among items that differentiated in different directions between the groups and other items that did not discriminate statistically between the groups. The evaluator compiled the new analysis in a supplementary report and sent it to the sponsor, along with the conclusion that the program clearly had succeeded in producing a set of important student outcomes. The evaluator also wrote a notable paper reflecting this

experience: *Needed: Instruments as Good as Our Eyes* (Brickell, 1976, 2011). We think his follow-up evaluation of the Ohio vocational education program was an early example of what Brinkerhoff (2003) later came to refer to as the Success Case Method, although statistical methods for evaluating "outliers" (which are usually considered noise or a nuisance by most statisticians) have existed for many decades.

Strengths

The Success Case Method is especially useful in ensuring that a program will be credited for whatever it has done well. When it is used as a formative evaluation approach, its principal strength is that it accelerates development by aiding early discovery of what is working and what is not. It is less often used as a summative method, but when it is, the approach's main strength is in ensuring that a program's positive points will be credited.

The orientation of the Success Case Method is to not "throw out the baby with the bathwater." As Brinkerhoff (2003) stated,

> Many evaluation approaches lead to overall "thumbs up or down" judgments, thus the few successes a program may be having get thrown out in the general bathwater of a larger initiative that is not working well. But, because the Success Case Method looks for success, no matter how small or infrequent, it helps new initiatives grow and become more successful. (p. x)

By comparing least successful instances to most successful instances, and by investigating as well the contextual factors and underlying causes that seem to contribute to success or a lack of it, the evaluator is often able to make useful suggestions for improving results. Even if a program is marginal or mostly poor in regard to quality and productiveness, it might be possible to find strengths on which the program could build (Brinkerhoff & Dressler, 2002). By identifying and understanding such strengths, the evaluator can discover why a program works and then help program leaders achieve more success (Brinkerhoff, 2005a). This also might help in deterring program funders from unjustifiably canceling programs that are partially succeeding or could be helped to succeed to a greater extent. Secondarily, discovering and documenting a program's successes can be instrumental in boosting program staff members' morale, giving them reasons to take pride in past accomplishments, and contributing concretely and publicly to a foundation of success on which they can build.

Weaknesses

The Success Case Method's main limitation is that the evaluator does not seek to produce a comprehensive assessment of an evaluand's merit and worth. Accordingly, it is best considered as an alternative approach that is especially useful in providing users with quick, reasonably rigorous, and typically low-cost responses to questions related to making an enterprise succeed. Compared with comprehensive assessments of a program's merit and worth, the Success Case Method is narrow in what it assesses and focuses mainly on short-term findings. This

narrowness is considered a weakness only if a study employing the Success Case Method is misrepresented as a comprehensive assessment of a program's merit and worth.

Approach 9: Outcome Evaluation as Value-Added Assessment

The classic example of value-added assessment involves collecting a standardized test score from each student in a school at the end of each of three or more successive school years, calculating each student's gain across the three or more scores, aggregating and analyzing the gain scores across all assessed students, and using the results to assess the school's effectiveness in improving its students' test scores. When this approach is applied to several schools in a school district, the resulting school-by-school results allow the evaluator to confidently compare the different schools in improving their students' test scores, irrespective of how high or low each school's students scored on the pretest. Systematic, recurrent outcome and value-added assessment, coupled with hierarchical gain score analysis, is a special case of the use of standardized testing to evaluate the effects of programs and policies. The emphasis is often on annual testing at all or a succession of grade levels to assess trends and partial out effects of an education system's different schools or other components.

Advance Organizers

Advance organizers in outcome evaluation employing value-added analysis are system-wide indicators of intended outcomes and a scheme for obtaining, classifying, and analyzing gain scores. The approach requires standardization of assessment data throughout a system. Questions to be addressed by outcome and value-added evaluations originate from governing bodies, policymakers, the system's professionals, and constituents. In reality, questions are often limited by the data available from tests regularly used by a state or school district.

Key variables in value-added assessment studies are units of measurement, including students, teachers, classrooms, schools, elementary and secondary levels of schools, curricular areas, school districts, and subgroups of school districts; selected standardized achievement tests; annual administrations of the tests; and student gain scores over a period of at least three years. With such elements in place, the investigator uses hierarchical, longitudinal analysis to identify achievement trends and associate the differential trends with the contributions of different schools, school districts, groups of districts, curricular areas, or teachers.

Purposes

The purposes of outcome and value-added assessment systems are to provide direction for policymaking, accountability to constituents, and feedback for improving programs and services. The intent is to determine what value each entity (school, school district, curricular area, or sometimes an individual teacher) is adding to the achievements of students served by an educational system and then report the results for policy, accountability, and improvement purposes. The main interest is in aggregates and trends across school years, not performance of individual students.

Sources of Questions

Key pressures for assessing educational outcomes emanate from oversight groups, such as state boards of education, school boards, and educational administrators that are under pressure to produce accountability reports. Such groups want and need studies that identify and assess academic achievement levels and trends in school districts and schools, and that link those assessments to the elements of the school district that may deserve recognition, require corrective attention, or need additional resources.

Questions

Basically, oversight bodies request answers to such questions as the following:

- What are the levels of and trends in achievement for each subgroup of school districts, each district, and each school in an education system?
- To what extent are particular programs adding value to students' achievements?
- What are the cross-year trends in outcomes?
- In what sectors of the system is the program working best, and where is it performing the most poorly?
- To what extent are program successes and failures associated with the system's groupings of grade levels (for example, primary, middle or junior high, and high school)?
- What are key, pervasive shortfalls—in particular, program objectives that require further study and attention?
- Which subgroups of districts, individual districts, schools, and teachers deserve commendation for their contributions to excellent levels of and trends in academic achievement?
- Which subgroups of districts, individual districts, schools, and teachers should be singled out and remediated or sanctioned because of deficient levels of and trends in academic achievement?
- To what extent do students sustain their pattern of test score gains as they move from one school building (say, an elementary school building) to another (a middle school building)?

Methods

A state education department may annually collect achievement test data from all students (at a succession of grade levels), as is the case in the Tennessee Value-Added Assessment System (see W. L. Sanders & Horn, 1994). The evaluator may analyze the data to look at contrasting gain score trends for different schools. Results may be broken out to make comparisons between curricular areas; teachers; elementary versus middle schools; or size and resource classifications of schools, districts, and areas of a state. What differentiates the approach from the typical standardized achievement testing program is the emphasis on sophisticated gain score and hierarchical analysis (for example, looking at students nested in classrooms nested in schools)—often referred to as hierarchical linear modeling (see Raudenbush & Bryk, 2002)

or multilevel modeling—of data to delineate effects of system components and identify which ones should be improved and which ones should be commended and reinforced. Otherwise, the two approaches have much in common.

Pioneers

Developers of the outcome evaluation as value-added assessment approach include W. L. Sanders and Horn (1994); Webster (1995); Webster, Mendro, and Almaguer (1994); and Tymms (1995).

Use Considerations

Outcome monitoring involving value-added assessment is probably most appropriate in well-financed state education departments and large school districts having strong support from policy groups, administrators, and service providers. The approach requires system-wide buy-in; politically effective leaders to continually explain and sell the program; annual testing at a succession of grade levels; a smoothly operating, dynamic, computerized baseline of relevant input and output information; highly skilled technicians to keep the computerized database up to date and make it run efficiently and accurately; a powerful computer system; complicated, large-scale statistical analysis; and high-level commitment to use the results for policy development, accountability, program evaluation, and improvement at all levels of the system.

Strengths

The central advantage of outcome monitoring involving value-added assessment is in the systematization and institutionalization of a database of outcomes that can be used over time and in a standardized way to study and find means to improve outcomes. This approach makes efficient use of standardized tests; is amenable to analysis of trends at state, district, school, and classroom levels; uses students as their own controls; and emphasizes that students at all ability levels should be helped to grow in knowledge and skills. The approach is conducive to using a standard of continuous progress across years for every student, as opposed to employing static cut scores. The latter, while prevalent in accountability programs, basically fail to take into account meaningful gains by low-achieving or high-achieving students, because such gains usually are far removed from the static cut score standards.

W. L. Sanders and Horn (1994) have shown that the use of static cut scores may produce a "shed pattern," in which students who began below the cut score standard make the greatest gains, whereas those who started above the cut score standard make little progress. Like the downward slope, from left to right, of a toolshed, the gains are greatest for previously low-scoring students and progressively lower for the higher achievers. This suggests that teachers may be concentrating mainly on getting students to the cut score standard but not beyond it, thus holding back the high achievers.

Figure 6.3, although not based on actual data, illustrates the shed pattern—in reference to the sloped roof of a toolshed—that W. L. Sanders and Horn (1994) have observed and reported

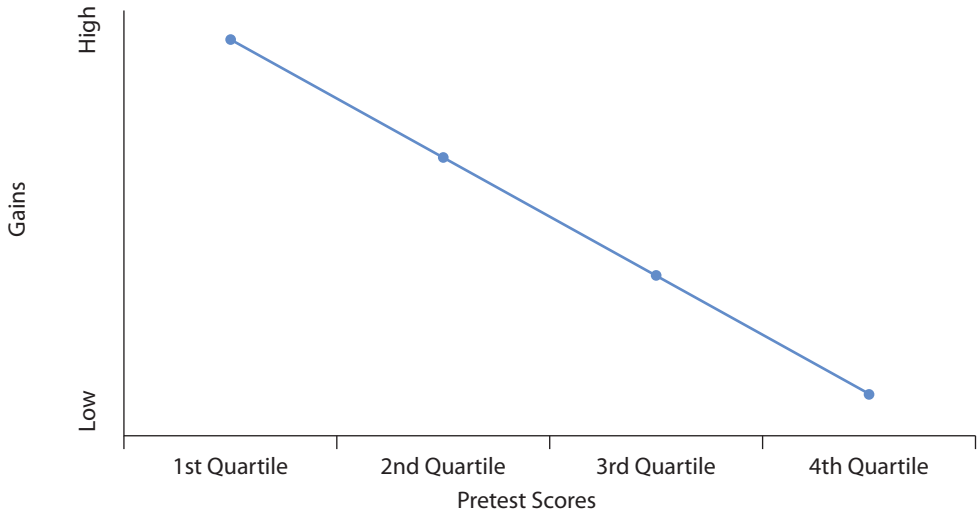

Figure 6.3 Hypothetical Shed Pattern of Student Gains over a Three-Year Period

in their applications of value-added assessment studies. The vertical dimension denotes gains, over a three-year period, in test scores for a sample of students. The horizontal dimension denotes lowest to highest quartiles of students in the sample based on average pretest scores for each quartile group. The slope in the figure is based on average three-year gains for each group. As shown, average gains are greatest for the quartile with the lowest pretest scores and are decreasingly lower for the second, third, and fourth quartiles. Thus there is an obvious negative correlation between pretest scores and posttest gain scores. Although one might argue that this pattern reflects the well-known phenomenon of regression to the mean, W. L. Sanders and Horn have posited that some of the regression very likely results from more intense instruction for low-scoring students and decreased instruction for those who already met or exceeded the cut score standard. This is especially likely in high-stakes testing programs, and we think W. L. Sanders and Horn have made a credible argument against the use of cut scores for judging students' school progress. It is also plausible that gains for high-scoring students are depressed by tests with relatively low ceilings.

Weaknesses

A major disadvantage of the outcome and value-added assessment approach is that it is politically volatile due to its use in identifying responsibility for successes and failures down to the levels of schools and teachers. It also is heavily reliant on quantitative information, such as that coming from standardized, multiple-choice achievement tests. Consequently, the complex and powerful analyses are based on a limited scope of outcome variables. Nevertheless, W. L. Sanders (1989) has argued that a strong body of evidence supports the use of well-constructed, standardized, multiple-choice achievement tests. Beyond the issue

of outcome measures, the approach does not provide in-depth documentation of program inputs and processes and makes little, if any, use of qualitative methods. Despite advancements in objective measurement and the employment of hierarchical linear modeling to determine effects of a system's organizational components and individual staff members, critics of the approach argue that causal factors are so complex that no measurement and analysis system can fairly fix responsibility for the academic progress of individual and collections of students to the level of teachers. Also, this book's first-named author's personal experience in interviewing educators in all of the schools in a Tennessee school district, subject to the statewide Tennessee value-added student assessment program (see W. L. Sanders and Horn [1994] for a complete description of the program), showed that none of the teachers, administrators, and counselors interviewed understood or trusted the fairness of this approach. That anecdotal finding may or may not reflect current stakeholder sentiments concerning the use of value-added assessment to judge teachers and schools, but it is a possible concern worth further investigation.

Approach 10: Experimental and Quasi-Experimental Studies

Using controlled experiments (and their many synonyms—*randomized controlled trials, randomized clinical trials, randomized experiments, true experiments,* and so on), program evaluators randomly assign recipients (such as students or groups of students or patients) or organizations (such as schools or hospitals) to experimental and control groups and then contrast outcomes after the experimental group has received a particular intervention and the control group has received no special treatment or some different treatment.

This type of study was quite prominent in program evaluations during the late 1960s and early 1970s, when there were federal requirements to assess the effectiveness of federally funded innovations in schools and social service organizations. In the 1980s and 1990s experimental program evaluations fell into disfavor and disuse. Apparent reasons for this decline were that educators, social workers, and other social service providers rarely can meet the required experimental conditions and assumptions. Recently, however, randomized experiments have returned to favor. Particularly influential on the renewed interest in using randomized experiments for program evaluation purposes has been the U.S. Department of Education Institute of Education Sciences (IES) and its constituents, which gave priority to randomized studies of educational interventions and innovations in 2004 (see Christie & Fleischer, 2010; Coryn, 2007a, 2011; Donaldson, Christie, & Mark, 2009). In 2010, however, IES modified its position to recognize other types of investigations (such as well-conducted quasi-experiments with statistical controls, certain single-subject designs, and case-control studies) as providing sufficient evidence of program effectiveness.

Advance Organizers

The advance organizers in experimental studies are problem statements, competing treatments, cause-and-effect hypotheses, investigatory questions, randomized treatment and comparison groups, defined dependent variables, and selected tools and procedures for obtaining dependent variable measures.

Purposes

The usual purpose of the controlled experiment is to determine causal relationships between specified independent and dependent variables, such as between a given method of instruction and student performance on a standardized test. The scientific rationale put forth to support the use of randomized experiments for program evaluation purposes is quite simple. Through randomization (with sufficiently large samples), all biases—measured and unmeasured—are distributed equally over treatment and control conditions (that is, the selection bias threat to internal validity is eliminated). Theoretically, therefore, the only possible explanation for outcome differences between groups is exposure to the treatment or intervention.

Sources of Questions

It is noteworthy that the sources of questions investigated in the experimental study are researchers, program developers, and policy figures, and not usually a program's constituents and staff. Requests, even pressures, for conducting randomized controlled experiments frequently come from oversight bodies, including federal funding agencies and the boards of charitable foundations. Such bodies often want to know the extent to which the programs they fund have produced positive outcomes. They often make clear that judgments of a program's success are unacceptable unless they are based on clear evidence that the funded program caused the measured effects.

Questions

The typical bottom-line question in an experimental study is, To what extent did a special treatment (for example, a new program) produce better outcomes than those observed for an alternative treatment? To effectively address such a question, the evaluator is expected to employ randomization and control of treatment conditions in determining unequivocal cause-and-effect relationships between independent and dependent variables. The independent variables are competing treatment conditions, and the dependent variables are the outcomes being sought.

Methods

The experimental method is employed through one of a range of experimental designs including two or more randomly assigned comparison groups as well as two or more treatment conditions. The array of available experimental designs includes those designated as posttest only, pretest-posttest, or Solomon four group, among others (see Campbell & Stanley, 1963).

Shown in Figure 6.4 is a flowchart presenting the typical progression of units through a randomized controlled experiment. Units can be individuals (for example, students) or clusters of units, such as schools, hospitals, communities, or charitable foundations (experiments involving clusters of units are known as group-randomized experiments). Once units are recruited and determined to be eligible, they are then randomly assigned to a condition

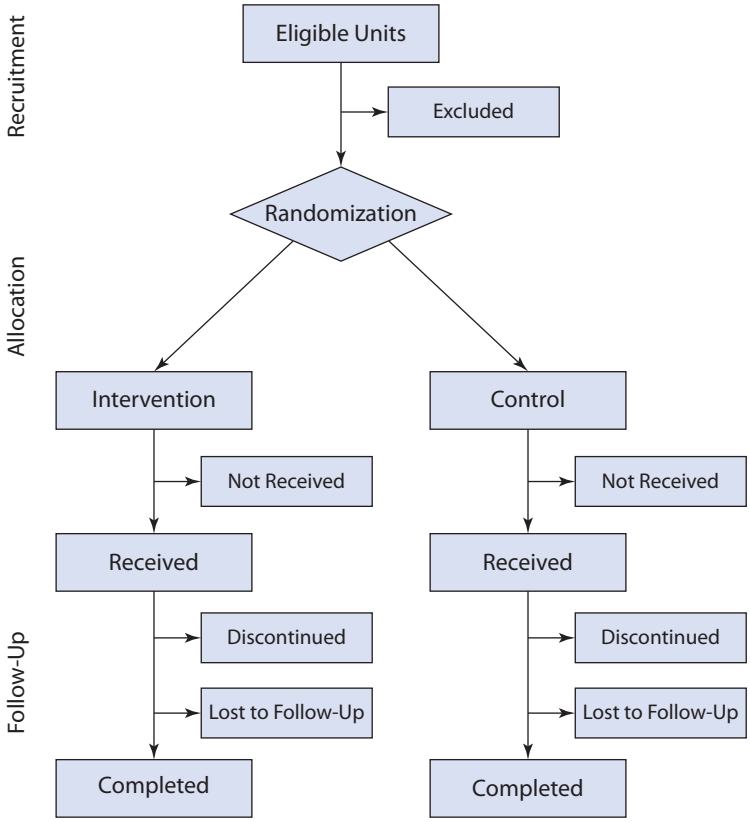

Figure 6.4 Flowchart of Units Through a Randomized Experiment

(typically to either a treatment or a control). Depending on the study's design and the nature of the intervention, a measure of the outcome of interest may be taken from members of each comparison group prior to the period of intervention to be studied. (Such measures are commonly referred to as pretests.) The comparison groups may include the subjects receiving an innovative treatment, the subjects undergoing current practice, or a group of subjects receiving no treatment whatsoever. Once groups have been exposed to the intervention or control, then posttesting occurs at some predetermined time when an effect is expected to have occurred or over several time points following a treatment (that is, repeated posttests).

Units that discontinue a treatment or drop out of a control condition, or that are lost to the follow-up measurement (in a process known as attrition), are often statistically retained in determining whether an intervention is effective through intention-to-treat (ITT) types of analysis (analysis based on the initial treatment intent, not on the treatment eventually administered). For purposes of ITT analysis, all who begin the treatment or intervention are considered to be part of the experiment, whether they finish or not, and those who did not complete are considered as a zero effect in the analysis. This at first seems counterintuitive. If an evaluator is trying to determine how effective a new drug might be, why would patients who

refused to take the drug be included? An analysis that includes only compliant participants has two major shortcomings:

- Groups defined by compliance are no longer randomized and are thus subject to biases.
- Groups defined by compliance may not represent the practical impact of the treatment.

When the analysis is conducted only on those who completed a program or treatment, known as treatment-on-the-treated (TOT) analysis, the randomized experiment essentially breaks down and becomes a quasi-experiment: those who remain or complete a program or treatment probably differ from those who started, thus introducing a selection bias, which randomization is designed to eliminate. Good practice would dictate that both ITT and TOT analyses be conducted, but this is rarely the case. More often than not, only TOT analysis is reported rather than ITT analysis (Kruse et al., 2002).

Although not shown in Figure 6.4, these types of studies can, and often do, include one or more pretests, more than two groups (for example, two treatment groups [one high dose, one low dose] and a control); various types of control groups (for example, no-attention control, wait-list control, or placebo control); multiple posttests; nonequivalent dependent variables (Coryn & Hobson, 2011); and other design elements intended to reduce different types of validity threats (see Shadish, Cook, & Campbell, 2002).

Pioneers

Pioneers in using experimental design to evaluate programs are Donald Campbell and Julian Stanley (1963), Lee Cronbach and Richard Snow (1969), E. F. Lindquist (1953), and Edward Suchman (1967). Others who have developed the methodology of experimentation substantially for program evaluation are Robert Boruch (1994, 2003); Gene Glass and Tom Maguire (1968); and David Wiley and Darrell Bock (1967), plus the authors referenced in this section.

Use Considerations

Evaluators should consider conducting a controlled experiment only when its required conditions and assumptions can be met. Often these conditions and assumptions include significant political influence, substantial funding, and widespread agreement—among the involved funders, service providers, and recipients—to submit to the requirements of the experiment. In addition, a true randomized, comparative experiment requires a stable program that will not have to be studied and modified during the evaluation; the ability to establish and sustain comparable program and control groups; the ability to keep the program and control conditions separate and uncontaminated; and the ability to obtain the needed criterion measures from all or at least representative samples of the members of the program and comparison groups. Evaluability assessment was developed as a particular methodology for determining the feasibility of moving ahead with an experiment that meets the necessary conditions (M. F. Smith, 1989; Wholey, 1995).

Due to some of the criticisms (such as those concerning withholding potentially effective treatments) and problems (such as implementation failures in field settings) associated with randomized experiments, developments and advances in quasi-experimental methods, largely advanced by Tom Cook (T. D. Cook & Campbell, 1979; Shadish et al., 2002), have been advocated as a means of producing credible causal evidence when randomized experiments are unethical or unfeasible. Quasi-experiments share many similarities with randomized experiments (for example, they too can use no-treatment control groups), except that units are not randomly assigned to treatment or control conditions and often consist of intact groups. Major developments in quasi-experimental methods include regression discontinuity designs (T. D. Cook, 2007), in which units can be assigned to conditions based on need using an assignment measure with a predetermined cut score to assign units to conditions (for example, low-performing students receive the intervention, and high-performing students do not), and interrupted time-series designs (Glass, Willson, & Gottman, 2008), both of which—if executed correctly—are capable of producing credible evidence of cause-and-effect relationships between an intervention and its outcomes.

Strengths

Controlled experiments have a number of advantages. They focus on results and not just on intentions or judgments. They provide strong methods for establishing relatively unequivocal causal relationships between treatment and outcome variables, something that can be especially significant when program effects are small but important. Moreover, because of the prevalent use and success of experiments in such fields as medicine and agriculture, the approach has widespread credibility.

In general, the experimental method is one that can make important contributions to program evaluation, as Nave, Miech, and Mosteller (2000) have demonstrated. However, as they and others (Spybrook, 2008; Spybrook & Raudenbush, 2009) have found, evaluators of educational and social programs rarely have executed sound and useful experiments, instead having conducted underpowered studies, assessed poorly executed interventions, or failed to recruit and retain all the units needed to validly assess treatment effects.

Weaknesses

The strengths of randomized experiments are offset by serious objections to experimenting on students and other human subjects. It is often considered unethical or even illegal to deprive control group members of the benefits of special funds for improving services. Likewise, many parents do not want schools or other organizations to experiment on their children by applying unproven interventions. Typically schools find it impractical and unreasonable to randomly assign students to treatments and to hold treatments constant throughout the study period. Furthermore, experimental studies provide a much narrower range of information than organizations often need to assess and strengthen their programs. On this point, experimental studies tend to provide terminal information that is not useful for guiding the development and improvement of programs and may in fact thwart ongoing modification of programs.

Approach 11: Cost Studies

Cost studies as applied to program evaluation involve a set of procedures used to estimate the costs of a program and to determine and judge what these investments returned in objectives achieved and broader social benefits.

Cost studies as commonly applied in program evaluation include cost-effectiveness, cost-benefit, and cost-utility analysis (Levin & McEwan, 2001; Yates, 1996). Straightforward documentation of a program's costs can also be extremely valuable to outsiders who might be interested in replicating a program, and documented costs are an essential precursor to the other types of cost studies.

Advance Organizers

Advance organizers for cost studies are associated with cost breakdowns for program inputs, outputs, and outcomes. Program input costs may be delineated by line item (personnel, travel, materials, equipment, communication, facilities, contracted services, overhead, and so on); program component; and year. Program outputs may be examined for immediate program results (for example, numbers of proposals submitted by a research laboratory to funding organizations or a precollege high school program's number of graduates). Program outcomes may be examined in terms of long-range benefits (for example, the number of proposals funded or number of precollege high school program graduates subsequently entering and graduating from a four-year college).

Purposes

The purposes of the main types of cost studies (that is, cost documentation as well as cost-effectiveness, cost-benefit, and cost-utility analysis) are to gain clear knowledge of what resources were invested, how they were invested, and with what immediate and long-term effects and, typically, to compare program costs to the costs of an alternative or standard.

Sources of Questions

In popular vernacular, cost analyses are used to determine a program's "bang for the buck." There is great interest in pursuing this line of inquiry. Policy boards, program planners, and taxpayers are especially interested in knowing whether program investments are paying off in terms of positive results that exceed or are at least as good as those produced by similar programs.

Questions

Particular cost-related questions are

- What are the costs associated with program inputs and other ingredients?
- What is the monetary value of program outcomes?

- What are pertinent computed cost ratios?
- How do the program's computed cost ratios compare to those of similar programs?
- Ultimately what is a program's level of productivity in economic terms?

Methods

Cost-effectiveness analysis may be done by itself. Such analysis involves comparing the relative costs and outcomes or effects of two or more courses of action in nonmonetary units (for example, tested math achievement as a measure of effectiveness). In all such analyses, time is a critical consideration, as costs, benefits, effectiveness, and other similar matters may differ over time.

In a cost-effectiveness ratio, the denominator is a unit of effectiveness (such as math achievement), and the numerator is the monetary value of all resources consumed to produce an outcome or effect, which can be expressed as

$$\frac{C}{E}$$

where C represents cost and E represents effectiveness. Cost-effectiveness analysis typically includes examining two or more programs' costs and successes in achieving the same objectives. A program could be judged superior on cost-effectiveness grounds if it had the same overall costs as similar programs but better outcomes. Or a program could be judged superior on cost-effectiveness grounds if it achieved the same objectives with fewer costs. Although cost-effectiveness analyses do not require conversion of outcomes into monetary terms, outcomes must be keyed to clear, measurable program objectives, and costs must account for inflation, depreciation, discounting, and uncertainties (Levin & McEwan, 2001). Also, such analyses cannot be used to assess the overall worth of a single intervention, and they are useful only for comparing two or more alternatives against a predetermined standard. The single intervention can, however, be compared to a predetermined standard or, formatively, examined to determine if costs are decreasing over time relative to effectiveness.

Cost-benefit analyses are used to look at costs associated with main effects and side effects, tangible and intangible outcomes, positive and negative outcomes, and short-term and long-term outcomes—both inside and outside a program. Frequently they also may involve breaking down costs by individual and by group of recipients. One may also estimate the costs of forgone opportunities and, sometimes, political costs.

Cost-benefit analysis typically builds on a cost analysis of program inputs and a cost-effectiveness analysis. The cost-benefit analysis is used to identify a broader range of outcomes than just those associated with program objectives. The investigator examines the relationship between the investment in a program and the extent of positive and negative impacts on the program's environment. In doing so, the investigator ascertains and places a monetary value on program inputs and each identified outcome. He or she identifies a program's benefit-cost ratios and compares these to similar ratios for competing programs. Ultimately, cost-benefit

studies lead to conclusions about the comparative benefits and costs of a program expressed as the ratio of benefits to costs:

$$\frac{B}{C}$$

where B is the number of monetary units of benefit for each unit of cost, C. If the benefit-cost ratio is greater than 1.00, it implies that benefits outweigh costs. Many types of program benefits, such as increases in family cohesion or a reduction in prejudice, are difficult to express in terms of monetary value, whereas others, such as reduced reliance on public assistance or increases in earnings associated with educational attainment, are more easily translated into monetary units.

In cost-utility analysis, various attributes of utility associated with different benefits are weighted according to stakeholder preferences using multiattribute utility theory, the direct method, or other decision theory approaches. After obtaining utility estimates for each alternative and its estimated cost, each cost estimate is divided to obtain a cost-utility ratio

$$\frac{C}{U}$$

where C represents the cost of each alternative and U represents its utility. The cost-utility ratios of the alternatives are rank-ordered from smallest to largest, with the smallest ratios indicating the alternatives that provide a given amount of utility at the lowest cost (Levin & McEwan, 2001). In cost-utility studies, utility is typically expressed as the preference for particular outcomes relative to others. For example, parents may believe that reading achievement is a more important outcome than math achievement and that therefore it should carry more weight in determining an educational program's cost-utility ratio.

Another type of cost study sometimes used for program evaluation purposes involves determining return on investment (Phillips, 2003), which can be expressed as

$$\frac{\text{Net program benefits}}{\text{Program costs}} \times 100$$

where net program benefits are the program benefits minus costs, with return on investment expressed as a percentage of a program's costs (the investment). Put another way, return on investment is the extent to which the benefits (outcomes) of a program exceed its costs (inputs). Such analyses are frequently used in corporate settings to estimate the benefits of training and other activities, such as increases in product sales or improved employee productivity.

Pioneers

Basically, the different types of cost studies have been developed by professionals in the fields of economics and financial accounting. Henry Levin has been a leading figure in bringing the methodology of cost analysis approaches to the evaluation field. Authoritative information on cost study approaches may be obtained by studying the writings of Kee (1995); H. M. Levin (1983; Levin & McEwan, 2001); Tsang (1997); and Yates (1996). In addition, Persaud (2007) has developed a useful checklist for cost analysis in program evaluation.

Use Considerations

Cost studies are potentially important in most program evaluations. Evaluators are advised to discuss this matter thoroughly with their clients, reach appropriate advance agreements on what should and can be done to obtain the needed cost information, and undertake as much cost-effectiveness, cost-benefit, and cost-utility analysis as can be done well and within reasonable costs. Because the requirements of such studies exceed the training of many evaluators, it is often necessary to team with or obtain assistance from experts in cost accounting, auditing, or economics. Even so, the basic requirements of cost analysis are easily learned and applied.

Strengths

The main strengths of cost studies reside in their explicit standards and methods that have been rigorously developed over the years in the fields of cost accounting and economics. Further strengths can be discerned by studying how economists such as Levin have applied cost analysis methodology to actual program evaluations.

Weaknesses

Documentation and analysis of costs are important but problematic considerations in program evaluations. For most evaluations, evaluators should at a minimum document the costs of program inputs and maintain a financial history of program expenditures. The main impediment is that program authorities often do not want anyone other than the appropriate accountants and auditors looking into their financial books. If a program's costs are to be studied, this must be provided for clearly in the initial contractual agreement covering the evaluation work. Performing a cost analysis can be feasible if the client formally sanctions the collection, analysis, and reporting of program costs; if there are clear, measurable program objectives; and if comparable cost information can be obtained from competing programs.

Unfortunately, it is usually hard to meet all of these conditions. Even more unfortunate is the fact that it is usually impractical to conduct a thorough cost-benefit analysis or cost-utility analysis. For one thing, such analyses must meet all the conditions of documentation of program costs and cost-effectiveness analysis. What is more, the evaluator must place monetary value on identified outcomes—anticipated and unexpected, short range and long range (in the case of cost-benefit analysis)—and, in the case of cost-utility analysis, obtain and analyze stakeholders' assignment of utility weights to different identified benefits. In many cases, the real value of benefits associated with human creativity or self-actualization is nearly impossible to estimate.

Approach 12: Connoisseurship and Criticism

The connoisseurship and criticism approach grew out of methods used in art criticism and literary criticism (Dewey, 1934), as well as those used to evaluate food, wine, and music (Stingley, 2010). It assumes that certain experts in a given substantive area are capable of

in-depth analysis and evaluation that could not be done in other ways. For example, a national survey of wine drinkers could produce information concerning their overall preferences for types of wine and particular vineyards, but it would not provide the detailed, creditable judgments of the qualities of particular wines that might be derived from a single connoisseur who has devoted a professional lifetime to the study and grading of wines and whose judgments are detailed and highly and widely respected.

In this type of study, an investigator with deep knowledge and experience in the realm of inquiry is the evaluation instrument. As Eisner (1998) noted, the word *connoisseurship* means "knowing." It involves the ability to see, not merely to look. To do this, he argued, one has to develop the ability to name and appreciate the different dimensions of situations and experiences, and the way they relate to one another. Connoisseurs have to be able to draw on, and make use of, a wide array of both experiences and information. They also have to be able to place their experiences and understandings in a wider context, and connect them with their personal values and commitments. Connoisseurship is something that needs to be worked at—but it is not a technical exercise. The bringing together of different elements into a whole involves artistry.

Eisner (1985) asserted, however, that educators need to become something more than connoisseurs—they need to become critics:

> If connoisseurship is the art of appreciation, criticism is the art of disclosure. Criticism, as Dewey pointed out in *Art as Experience,* has at its end the re-education of perception . . . The task of the critic is to help us to see . . . [and] connoisseurship provides criticism with its subject matter. Connoisseurship is private, but criticism is public. Connoisseurs simply need to appreciate what they encounter. Critics, however, must render these qualities vivid by the artful use of critical disclosure. (pp. 92–93)

Advance Organizers

The advance organizers for a study based on connoisseurship and conveyance of criticism-based descriptions and judgments are an evaluator's special expertise, sensitivities, tacit knowledge, and refined capability to portray and communicate.

Purposes

Such a study's purpose is to describe, critically appraise, and illuminate a particular program's characteristics, strengths, and weaknesses.

Sources of Questions

Evaluation questions addressed by connoisseurship and criticism evaluations are determined by expert evaluators—critics and authorities who have undertaken the evaluation.

Questions

Among the major questions the connoisseur-critic evaluator can be expected to ask are these: What are a program's essence and salient characteristics? What merits and demerits distinguish a particular program from others of the same general kind?

Methods

The methodology of connoisseurship and criticism includes the critic's systematic use of his or her perceptual sensitivities, past experiences, refined insights, and ability to communicate his or her assessments. In some areas of evaluation practice this approach is referred to as sensory evaluation (Barrett, 2011). An evaluator's judgments are conveyed in vivid terms to help the audience appreciate and understand all of the program's nuances.

Pioneers

Elliott Eisner (1975, 1983, 1985, 1998, 2004) pioneered this strategy in education. A dozen or more of Eisner's students have conducted research and development on the connoisseurship and criticism approach, including Vallance (1973) and Flinders (Flinders & Eisner, 2000).

Use Considerations

This approach obviously depends on the chosen expert's qualifications. It also requires an audience that has confidence in, and is willing to accept and use a report from, an evaluator who employs an approach based on connoisseurship and criticism.

Strengths

The main advantage of an effectively conducted connoisseurship and criticism study is that it exploits the particular expertise and finely developed insights of a specialist who has devoted much time and effort to the study of a precise area. Such an individual can provide an array of detailed information that an audience can then use to form a more insightful analysis than otherwise might be possible.

Weaknesses

The approach's disadvantage is that it is dependent on the expertise and qualifications of the particular expert doing the evaluation, leaving room for much subjectivity. It can be difficult for a program's stakeholders to agree on the one or few connoisseur-critic evaluators they would all trust to deliver an acceptable evaluation. After the report has been delivered, its possibly controversial message may be rejected on the grounds that it was not objectively determined and delivered. Thus, this approach may be highly subject to political attack.

Approach 13: Theory-Based Evaluation

Program evaluations based on a program theory often begin with either (1) a well-developed and validated theory of how programs of a certain type within similar settings operate to produce outcomes, or (2) an initial stage to approximate such a theory within the context of a particular program evaluation. The former condition is much more reflective of the implicit promises inherent in a theory-based evaluation, because the existence of a sound theory means that a substantial body of theoretical development has produced and tested a coherent set of conceptual, hypothetical, and pragmatic principles, as well as associated instruments to guide inquiry. The theory can then help a program evaluator decide what questions; indicators (that is, manifest variables); and linkages (assumed to be causal) between and among program elements should be used to evaluate a program covered by the theory. Often, such a theory is presented in the form of a linear model of how a program is anticipated to produce a certain effect (see Figure 6.5).

Given recent criticisms of the linear program theory model, some systems theorists (for example, Morell, 2010; Patton, 2010; Williams & Imam, 2007) have argued that nonlinear, ecological, open, adaptive, and complex theories of a program often are more appropriate approximations of true program theories and are more consistent with reality than linear theories (see Figure 6.6). Even so, and although theory-driven and systems theories and approaches to evaluation have long been advocated, we do not recommend their use as frameworks for comprehensive evaluation. These two theory-oriented approaches lack sufficient validation for use in program evaluation; both are short on well-developed and tested methods; and even the distinction between them is unclear (also see Scriven, 2005e).

Some theories have been used more or less successfully to evaluate programs, and this gives the approach some measure of viability. For example, health education and behavior change programs are sometimes founded on theoretical frameworks, such as the health belief model (Becker, 1974; Janz & Becker, 1984; Mullen, Hersey, & Iverson, 1987). Other examples are the PRECEDE-PROCEED model for health promotion planning and evaluation (Green & Kreuter, 1991); Bandura's social cognitive theory (1977); the "stages of change" theory of Prochaska and DiClemente (1992); and Peters and Waterman's theory (1982) of successful organizations. When such frameworks exist, their use probably can enhance a program's effectiveness and provide a credible structure for evaluating a program's functioning. Unfortunately, few program areas are buttressed by well-articulated and tested theories, and alternative theories

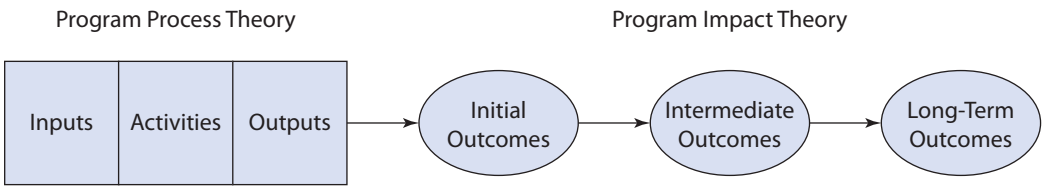

Figure 6.5 Linear Program Theory Model

Source: Donaldson, S. I. (2007). *Program theory–driven evaluation science*. Mahwah, NJ: Lawrence Erlbaum Associates, 25.

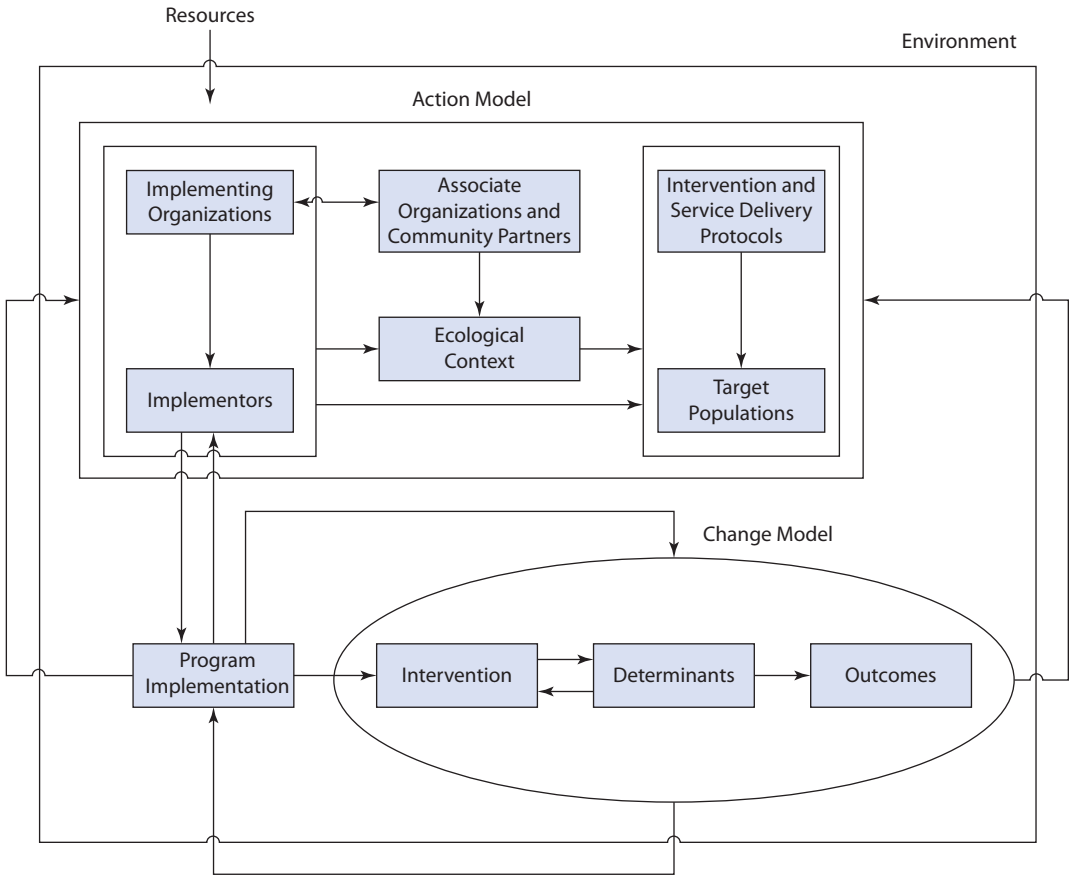

Figure 6.6 Nonlinear Program Theory Model

Source: Chen, H. T. (2005). *Practical program evaluation: Assessing and improving planning, implementation, and effectiveness*. Thousand Oaks, CA: Sage, 31.

(whether derived from formal theories or from program stakeholders) are not frequently tested (Coryn, 2009; Coryn, Noakes, Westine, & Schröter, 2011).

In most theory-based evaluations, the evaluator begins by setting out to develop a theory (typically in the form of a logic model or similar conceptualization) that appropriately could be used to guide a particular program evaluation. As will be discussed later in this characterization, linking theory development efforts to program evaluations is problematic and potentially counterproductive. In any case, let us look at how the evaluator incorporates theory into the planning and conduct of program evaluations.

Advance Organizers

The point of a theory development or selection effort is to identify advance organizers (for example, in the form of a measurement model) to guide the evaluation's collection and analysis of information. Essentially, advance organizers are the mechanisms by which program activities

are understood to produce or contribute to program outcomes, along with the appropriate description of context, specification of independent and dependent variables, and portrayal of key linkages.

Purposes

The main purposes of a theory-based evaluation of a program are to determine the extent to which the program of interest is theoretically sound, to understand why it is succeeding or failing, and to provide direction for program improvement (Rogers, 2000).

In summing up and reporting on a theory-based evaluation, the evaluator seeks to present conclusions, pro or con, on the program's theoretical soundness, its operation in accordance with an appropriate theory, its production of expected outcomes, its confirmation of hypothesized causal linkages, its execution as planned, any modifications in aims or procedures, and its worthiness for continuation and/or dissemination.

Sources of Questions

Questions for theory-based evaluations pertain to and are derived from the guiding theory.

Questions

Example study questions include the following:

- Is the program grounded in an appropriate, well-articulated, and validated theory?
- Is the employed theory reflective of sound research?
- Are the program's targeted recipients, design, operation, and intended outcomes consistent with the guiding theory?
- How well does the program address and serve the targeted recipients' full range of pertinent needs?
- If the program is consistent with the guiding theory, are the expected results being achieved?
- Are program inputs and operations producing outcomes in the ways the theory predicts?
- What changes in the program's design or implementation might produce better outcomes?
- What elements of the program are essential for successful replication?

The nature of these questions suggests that the success of the theory-based approach is dependent on a foundation of sound theory development and validation. This, of course, entails sound conceptualization of at least a context-dependent theory, formulation and rigorous testing of hypotheses derived from the theory, development of guidelines for practical implementation of the theory based on extensive field trials, development of valid instruments for assessing key aspects of the theory, and independent assessment of the theory.

Methods

The main element of a theory-based evaluation of a program typically is a model of the program's logic (Funnel & Rogers, 2011). This may be a detailed flowchart of how inputs are thought to produce intended outcomes, as previously shown in Figures 6.5 and 6.6. The foundational element may also be a grounded theory, such as those advocated by B. G. Glaser and Strauss (1967). The analysis involved in the flowchart approach is typically an armchair theorizing process involving evaluators and persons who are supposed to know how the program is expected to operate and produce results (in other words, leading to a stakeholder model of the program theory). They discuss, scheme, discuss some more, discuss further, and finally produce networks in varying degrees of detail of what is involved in making the program work and how the various elements are linked to produce desired outcomes. The more demanding grounded theory approach requires a systematic, empirical process of observing events or analyzing materials drawn from operating programs, followed by an extensive modeling process—in other words, observation of the program in operation (Coryn, Noakes, et al., 2011).

Such an approach using grounded theory is reminiscent of the intermittent theorizing and testing process that Wilbur and Orville Wright painstakingly employed, over a period of years in their bicycle shop and at Kitty Hawk, to ultimately develop and demonstrate—after much trial and error—the first successful flying machine. In retrospect, one might see the key advance organizers of their effort as a homemade wind tunnel; wing, propeller, and engine designs; proper fabric; controls; and the landing apparatus. Even today, aeronautical engineers credit the Wright brothers' calculations, detailed notes, and planes as the basis for much of current theory that guides development and testing of modern air machines. We cite this example both to highlight the value of sound theories for use in guiding evaluations and to stress that the demands of sound theory development far exceed the time, resources, and expertise available to evaluators in most program evaluations.

Pioneers

Pioneers in applying theory development or theory selection to program evaluation include B. G. Glaser and Strauss (1967) and Weiss (1972, 1995, 1997a, 1997b, 1998, 2000, 2004a, 2004b). Other developers of the approach include Bickman (1987, 1996); Chen (1989, 1990, 1994, 2005); Donaldson (2007); Rossi (Chen & Rossi, 1983, 1987, 1992); and Rogers (2000; Funnel & Rogers, 2011).

Use Considerations

In their systematic review of theory-driven evaluation practice, Coryn, Noakes, et al. (2011) found that few practitioners who claimed to be using a theory-driven evaluation approach actually applied all of the core principles and subprinciples of theory-based evaluation (see Exhibit 6.1). In fact, a majority of practitioners applied fewer than half of the theory-based evaluation approach's principles in practice.

Exhibit 6.1 CORE PRINCIPLES AND SUBPRINCIPLES OF THEORY-DRIVEN EVALUATION

1. Theory-driven evaluations/evaluators should formulate a plausible program theory
   a. Formulate program theory from existing theory and research (e.g., social science theory)
   b. Formulate program theory from implicit theory (e.g., stakeholder theory)
   c. Formulate program theory from observation of the program in operation/exploratory research (e.g., emergent theory)
   d. Formulate program theory from a combination of any of the above (i.e., mixed/integrated theory)
2. Theory-driven evaluations/evaluators should formulate and prioritize evaluation questions around a program theory
   a. Formulate evaluation questions around program theory
   b. Prioritize evaluation questions
3. Program theory should be used to guide planning, design, and execution of the evaluation under consideration of relevant contingencies
   a. Design, plan, and conduct evaluation around a plausible program theory
   b. Design, plan, and conduct evaluation considering relevant contingencies (e.g., time, budget, use)
   c. Determine whether evaluation is to be tailored (i.e., only part of the program theory) or comprehensive
4. Theory-driven evaluations/evaluators should measure constructs postulated in program theory
   a. Measure process constructs postulated in program theory
   b. Measure outcome constructs postulated in program theory
   c. Measure contextual constructs postulated in program theory
5. Theory-driven evaluations/evaluators should identify breakdowns and side effects, determine program effectiveness (or efficacy), and explain cause-and-effect associations between theoretical constructs
   a. Identify breakdowns, if they exist (e.g., poor implementation, unsuitable context, theory failure)
   b. Identify anticipated (and unanticipated), unintended outcomes (both positive and negative) not postulated by program theory
   c. Describe cause-and-effect associations between theoretical constructs (i.e., causal description)
   d. Explain cause-and-effect associations between theoretical constructs (i.e., causal explanation)
      i. Explain differences in direction and/or strength of relationship between program and outcomes attributable to moderating factors/variables

ii. Explain the extent to which one construct (e.g., intermediate outcome) accounts for/mediates the relationship between other constructs

Source: Coryn, C.L.S., Noakes, L. A., Westine, C. D., & Schröter, D. C. (2011). A systematic review of theory-driven evaluation practice from 1990 to 2009. *American Journal of Evaluation, 32,* 205.

Strengths

In any program evaluation assignment, it is reasonable for the evaluator to examine the extent to which program plans and operations are grounded in an appropriate theory or model. It can also be useful to engage in a modicum of effort to network the program and thereby seek out key variables and linkages. Modest attempts to create program models—labeled as such—can be useful for identifying measurement variables, so long as the evaluator does not spend too much time on this, and so long as a model is not considered to be a fixed or validated theory. Fortunately, the published methods of evaluation provide clear, useful direction for developing logic models and other schemes for representing and analyzing the interplay and timing of a program's procedures, milestones, and intended outcomes. In the enviable but rare situation where a relevant, validated theory exists, an evaluator can beneficially apply it in structuring the evaluation and analyzing findings.

Weaknesses

If a relevant, defensible theory of the program's logic does not exist, evaluators need not develop one. In fact, if they attempt to do so, they will incur many threats to the evaluation's success. Rather than evaluating a program and its underlying logic, evaluators might usurp the program staff's responsibility for program design. They might do a poor job of theory development, given limitations on time and resources to develop and test an appropriate theory. They might incur the conflict of interest associated with having to evaluate the theory they developed. They might pass off an unvalidated model of the program as a theory, when it meets almost none of the requirements of a sound theory. They might bog down the evaluation by expending too much effort to develop a theory. They might also focus attention on a theory developed early in a program and later discover that the program has evolved to be a quite different enterprise from what was theorized at the outset, in which case the initial theory could become a procrustean bed for both the program and the program evaluation.

Unfortunately, not many program areas in education and the social sciences are grounded in sound theories. Moreover, evaluators wanting to employ a theory-based evaluation do not often find it feasible to wait for the program staff to conduct the full range of theory development and validation steps before proceeding with the evaluation and still get the evaluation done effectively, on time, and within budget. Thus, in proposing to conduct a theory-based evaluation, evaluators often seem to promise much more than they can deliver. Overall, there is not much to recommend theory-driven program evaluation because

few validated theories exist for use in evaluating programs, doing theory-driven evaluation correctly is usually not feasible, the evaluator is not the right party to develop a desired program theory, and failed or misrepresented theory-driven evaluation attempts can be highly counterproductive.

Approach 14: Meta-Analysis

Research reviews, research syntheses, and meta-analysis are approaches whereby evaluators synthesize and draw conclusions from information provided by a set of similar studies. Such approaches are premised on the assumption that although individual studies provide only limited information about an intervention's effectiveness, each can contribute to a larger knowledge base of information on the intervention's effectiveness.

Examples of research reviews are the literature reviews found in doctoral dissertations, wherein the doctoral student identifies, summarizes, and analyzes previous studies that addressed questions akin to those being investigated in the dissertation. Here the student seeks to learn and convey not only the range and central tendencies of outcomes reported across these studies but also the range and attributes of needs and problems studied and the inquiry methods that were employed. Similarly, research syntheses involve compiling and analyzing research findings from all relevant studies to address particular questions. For example, a synthesis panel of expert cardiologists might collect, summarize, and analyze studies of medical practices, such as the use of angiography to diagnose cardiovascular disorders. The panel might subsequently issue a best practices report on the use of this procedure. In conducting a meta-analysis, an investigator states a hypothesis about the relative merits of certain alternative treatment conditions, collects reports of studies in which the treatment conditions of interest have been experimentally compared, aggregates the reported outcomes for the treatment and comparison groups, and conducts appropriate statistical tests to determine the significance of differences between aggregated treatment group outcomes and aggregated comparison group outcomes. The investigator concentrates almost exclusively on measures of intended outcomes, but may also look for pervasive patterns of side effects. The remainder of this section is focused on our characterization and assessment of the meta-analysis approach.

Purposes

Fundamentally, the purpose of a meta-analysis is to collect, summarize, analyze, and draw conclusions about an intervention's effects, as discernible from multiple credible studies of the intervention. Meta-analyses are employed to aggregate and assess overall results from multiple comparative, experimental studies. The selected studies may include both true randomized, comparative experiments and nonrandomized quasi-experiments. The principal ideology underlying meta-analysis (and other data synthesis approaches) is that an evaluation should not necessarily be viewed in isolation, but rather may be investigated as one of a set of tests of a program or intervention to explore collective results across variations in persons, treatments, outcomes, contexts, and other variables (Chalmers, 2007).

Borenstein, Hedges, Higgins, and Rothstein (2009) noted in the preface to their book, *Introduction to Meta-Analysis,* that meta-analysis studies can be, and often are, an ethical imperative for determining whether a treatment or intervention is safe and effective. They pointed out that through meta-analyses, long-accepted practices, such as Benjamin Spock's published advice to have babies sleep on their stomach, have been shown through compilations and analysis of relevant evidence to be harmful and in many instances fatal. As is now well established through summarization and assessment of actuarial data, the practice of having babies sleep on their stomach has been found to be strongly associated with a large number of crib deaths.

Advance Organizers

Advance organizers of meta-analyses are

- A sufficient number of studies of similar treatments or programs to permit a reliable and valid synthesis
- Important policy-oriented questions that can be addressed through analysis of findings from the set of studies
- Comparable, defensible outcome data from all studies in the set
- Defensible data collection and analysis designs that guided all studies in the set
- Relevant unpublished reports as well as published reports (to militate against publisher bias associated with publishing studies showing significant differences)
- A sound meta-analysis design for summarizing outcome data, computing effect sizes, analyzing the statistical significance of differences between aggregated experimental and comparison group outcomes, and gauging the practical significance of any observed differences between experimental and comparison group outcomes

Sources of Questions

Increasingly, policy groups have advocated the conduct of studies employing meta-analysis, seeing such studies as a vital source of information for making informed policy and practice decisions (for example, Cooper, 2010; Cooper, Hedges, & Valentine, 2009; Hunter & Schmidt, 2004; Lipsey & Wilson, 2001).

Questions

The fundamental questions that policy groups hope meta-analysis studies will help answer are

- What is the average effect of a given intervention across a set of studies that were completed to determine the intervention's effects compared to those of some other treatment condition?
- How significant in both statistical and practical terms are any observed differences between average effects of the different treatment conditions?

Methods

An evaluator taking a meta-analysis approach employs statistical methods for integrating results from multiple studies of similar programs or interventions and determining the statistical significance of differences between aggregated treatment outcomes and aggregated control group outcomes. In a meta-analysis, the unit of analysis is research reports rather than subjects (such as people).

A hypothetical example of results derived from a typical meta-analysis is shown in Figure 6.7 (although such studies often include analysis of one or more moderator variables—that is, factors or variables over which treatment effects may vary, such as gender, study design, treatment dosage, or variation in program implementation). In the example, the question addressed simply is whether a high dose (the experimental condition) or standard dose (the control condition) of statins reduces the risk of death due to myocardial infarction. In the figure, each study's effect size is represented as a risk ratio. A risk ratio of 1.00 represents no difference between two treatments (high dose and low dose)—that is, an equal 1:1 risk. A risk ratio less than 1.00 represents an effect favoring the high-dose treatment—that is, a reduced risk of death—and is denoted as a square. The square's relative size indicates the study's weight in the computation of the weighted average, or summary, effect size (denoted by a diamond). The diamond represents the bottom-line synthesis of study results. The horizontal line running through each study's estimated effect size represents each study's precision (that is, sampling error) in the form of a 95 percent confidence interval. Less precise studies (such as

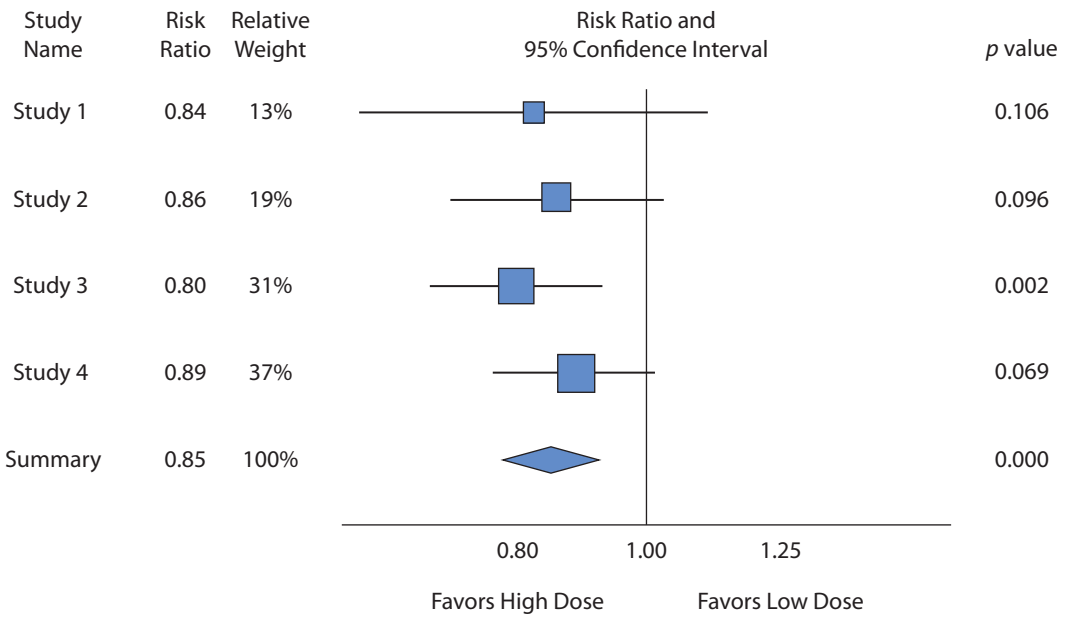

Figure 6.7 Hypothetical Meta-Analysis Forest Plot

Source: Adapted from Borenstein, M., Hedges, L. V., Higgins, J.P.T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, West Sussex, UK: Wiley.

Study 1) carry less weight in the summary effect than more precise studies (such as Study 4) in random-effects models. The 95 percent confidence interval—or precision—for the summary effect, accounting for the variability of each individual study, is reflected in the width of the diamond (in the example shown in Figure 6.7, extending from 0.79 to 0.92), which is an estimate of the intervention's true effect.

In general, as indicated by the diamond, the data summarized in this chart indicate that high-dose statins proved substantially better than low-dose statins in reducing the risk of death from a heart attack. All four studies supported this finding. Only Study 1 had a possibly worrisome level of measurement error as indicated by its wide confidence interval and *p* value of 0.106. However, as shown in the chart by the relative smallness of the box representing Study 1 and its 13 percent contribution to the overall effect size, this study had little influence on the determination of the bottom-line effect size and interpretation.

Pioneers

The origins of the meta-analysis approach can be traced back over one hundred years (for example, to Pearson's review (1904) of evidence on typhoid vaccines). In 1976, Glass introduced his rendition of the meta-analysis procedure. Glass's seminal contribution spawned extensive development, adaptation, and application of the methodology of meta-analysis.

Use Considerations

There have emerged very useful sources of research findings that can be used in meta-analysis studies. In recent years numerous evidence-based research repositories, such as the Cochrane Collaboration, Campbell Collaboration, and What Works Clearinghouse, among many others, have surfaced to provide cumulative evidence on what works and what does not across an increasingly diverse set of programs and interventions in health and medicine, education, and nearly all forms of human and social services (Coryn, Tarsilla, & Hobson, 2010). Presently, nearly all federally funded and non–federally funded research repositories rely almost exclusively on the results of meta-analyses as the primary evidentiary basis for informing "best" policies and "best" practices.

Bias in meta-analysis studies may be present when included studies are overly reflective of studies that found significant differences between outcomes for treatment and control groups. Such bias can result from including only published reports, because journals often do not report studies that found no significant differences. One means of reducing the likelihood of such publication bias is to include data from both published and unpublished reports.

Strengths

Meta-analysis studies clearly have legitimate uses within the sphere of sound evaluation services. Investigators exploit the potential lessons from completed studies in a given terrain by carefully selecting similar studies, aggregating their outcome data, and determining whether a treatment condition repeated across many studies produced better results than an alternative

treatment. Meta-analysis studies clearly have proved their value in addressing important policy questions. Moreover, the developers of the meta-analysis approach have contributed rigorous methods for applying the approach. Many investigators have found these methods to be feasible for effective application. Meta-analysis rates high on feasibility because investigators focus on variables that have been measured well in a set of juried studies, because proponents of the approach stress the need to meet standards of technical quality, because investigators examine important though narrow questions, and because there are many audiences for such focused studies.

Weaknesses

The meta-analysis approach mainly is limited by a singular focus on program outcomes rather than on a comprehensive assessment of all important dimensions of a program's merit and worth. Also, its scope is limited to the data that have been gathered in past studies. Another limitation is that the approach is not responsive to a program's evaluation needs during the program's execution. As with other quasi-evaluation approaches, meta-analysis must not be represented as a procedure for fully evaluating a program's value. Although measured outcomes are a vital part of comprehensive program evaluation, there are many other variables that must be examined in the course of fully assessing a program's merit and worth. Clearly, meta-analyses are useful in their own right, and such studies can also be included as a valuable component of larger, comprehensive program evaluations.

Summary

This chapter has

- Identified a class of quasi-evaluation approaches, including objectives-based studies, the Success Case Method, value-added assessment, experimental and quasi-experimental studies, cost studies, connoisseurship and criticism, theory-based evaluation, and meta-analysis
- Explained that some approaches are labeled as quasi-evaluation because typically they guide studies that are valuable in the focused evaluative feedback they yield but that typically are too narrow to meet the requirements of a comprehensive assessment of a program's merit and worth
- Identified functions, strengths, and weaknesses of quasi-evaluation approaches
- Identified, characterized, and assessed a sample of quasi-evaluation approaches

Quasi-evaluation approaches are highly valuable to evaluation clients in effectively addressing selected, specific questions and producing timely, focused reports. Studies based on these approaches are unlikely, however, to fully address an evaluation's fundamental requirement to assess a program's merit and worth.

The main caution is that narrow-scope, quasi-evaluation studies should not be uncritically equated with evaluations that fully assess an evaluand's merit and worth. That said, it is in every evaluator's best interest to develop functional levels of understanding of the approaches reviewed in this chapter, including their functions, strengths, and weaknesses. Working knowledge of these approaches can only enhance an evaluator's repertoire and ability to respond usefully and appropriately to requests for different types and levels of evaluation services. It is clear that clients often request certain narrow evaluation services that legitimately may be delivered using an appropriate quasi-evaluation approach.

REVIEW QUESTIONS

1. Respond to the charge that a quasi-evaluation often but not always is too narrow to be considered as a comprehensive assessment of a program's merit and worth.
2. Write a definition for each of the following evaluation approaches:
   a. Objectives-based evaluation
   b. Outcome evaluation as value-added assessment
   c. Connoisseurship and criticism
   d. The Success Case Method
   e. Meta-analysis

   Check your definitions against those provided in the glossary at the back of this book.
3. In designing an experimental study to compare the effectiveness of a hospital's three new neighborhood-based urgent care centers with that of its long-standing centralized emergency room, what would you include as advance organizers for the study?
4. Suppose that (a) an elementary school's After-School Study and Tutoring Program (ASSTP) is receiving criticism from school board members for being too costly and not really essential, and (b) you have accepted the ASSTP parent advisory board's request that you perform a Success Case Method study as a means of possibly preventing the program's termination. List examples of questions you probably would need to address in conducting the study.
5. In explaining the theory-based approach to a potential evaluation client, what would you list as (a) this approach's strengths and weaknesses, and (b) the prerequisites for applying it successfully?
6. Those subjected to outcome evaluation as value-added assessment have often raised objections to its use. What are the main objections? In the face of such objections, what might you cite as the approach's offsetting strengths and the reasons for its continued use? Also, what would you list as the necessary preconditions for applying the approach?
7. Identify some program with which you are familiar—such as a school's cafeteria food service program—and assume you have agreed to evaluate the program's cost-effectiveness. Then (a) state a feasible purpose for this study, (b) list needed advance organizers for the study,

(c) list potential sources of cost-related questions, (d) list example questions for the study, (e) summarize the procedures you would apply, and (f) in general terms, outline the report you would produce.

8. If a program evaluation is to be theoretically based, (a) What are the two (alternative) beginning conditions? and (b) What are the advantages and disadvantages of each of these beginning conditions?
9. Identify a program or practice—such as a school district's use of mandatory busing of students to fully integrate all of the district's schools—that you see as amenable to meta-analysis. Then (a) state a purpose for this study; (b) list pertinent advance organizers for the study; (c) list potential sources of questions that would guide the study; (d) list example questions for the study; (e) summarize the data collection and analysis procedures you would apply; (f) in general terms, outline the report you would produce; and (g) list what you see as the primary advantages and disadvantages of meta-analysis as an approach for evaluating the program.
10. What are the similarities and differences between the cost documentation, cost-effectiveness analysis, cost-benefit analysis, and cost-utility analysis approaches?

Group Exercises

This has been a lengthy chapter containing considerable information about the nature of eight quasi-evaluation approaches, all of which often are too narrow to fully assess a program's merit and worth. Nevertheless, most are commonly used, which is unlikely to change. We hope that your discussions will sharpen your views about some of the salient features of these quasi-evaluation approaches.

Exercise 1

Appoint a member of your group to chair this exercise. Divide the group into sections 1 and 2. Each individual in section 1 selects a different quasi-evaluation approach and states opinions, for all to hear, as to why it is useful to organizations willing to accept it. Members of section 2 listen to and take notes on the oral reports from members of the first section. Members of section 2 then outline perceived weaknesses of each reviewed approach. Subsequently, the chair leads a discussion of the strengths, weaknesses, and potential utility of the reviewed quasi-evaluation approaches. (The session chair will need to be a stern adjudicator in controlling this discussion!)

Exercise 2

Divide your group into four subgroups. Each subgroup should discuss and pass judgment on one of the following statements concerned with the theory-based approach, such that all statements

are addressed. Then convene the entire group and have one member of each subgroup give a five-minute report of his or her subgroup's reactions to the assigned statement. After all four reports have been presented, discuss as a whole group the merits of the theory-based approach. Following are the four statements to be examined and discussed:

a. Sound theory-based evaluations of programs are seldom feasible, because few validated program theories are readily available and applicable to given program evaluation assignments.

b. Conducting theory development in the course of planning a program evaluation carries such hazards as producing a poor theory, sapping evaluation resources for the theory development effort, impeding the program evaluation's progress, and putting the evaluator in the compromising position of having to evaluate the theory he or she developed.

c. In contracting for a program evaluation, the client should provide a long enough timeline and sufficient resources to permit the evaluator to develop and validate a sound, relevant theory before proceeding with the needed data collection and analysis stages. Sound evaluations aren't easy or cheap, and the client should take the long view and be ready to invest whatever resources and time are required to reach sound evaluative conclusions.

d. The client need not worry about the evaluator's conflict of interest in evaluating the theory he or she developed, because evaluation is inevitably a subjective process anyway.

Exercise 3

Discuss the pros and cons of using the value-added assessment approach to evaluate teachers.

Note

1. The quasi-evaluation approach labeled "case study evaluation" is not summarized and analyzed in this chapter, but it is defined extensively in Chapter 12.

Suggested Supplemental Readings

Borenstein, M., Hedges, L. V., Higgins, J.P.T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, West Sussex, UK: Wiley.

Boruch, R. F. (1998). Randomized controlled experiments for evaluation and planning. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 161–192). Thousand Oaks, CA: Sage.

Boruch, R. F. (2003). Randomized field trials in education. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 107–124). Norwell, MA: Kluwer.

Brinkerhoff, R. O. (2003). *The Success Case Method: Find out quickly what's working and what's not*. San Francisco, CA: Berrett-Koehler.

Brinkerhoff, R. O. (2005). The Success Case Method: A strategic evaluation approach to increasing the value and effect of training. *Advances in Developing Human Resources*, *7*, 86–101.

Brinkerhoff, R. O. (2006). *Telling training's story: Evaluation made simple, credible, and effective*. San Francisco, CA: Berrett-Koehler.

Chen, H. T. (1990). *Theory-driven evaluations*. Thousand Oaks, CA: Sage.

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Applied Social Research Methods Series, Vol. 2. Thousand Oaks, CA: Sage.

Coryn, C.L.S., Noakes, L. A., Westine, C. D., & Schröter, D. C. (2011). A systematic review of theory-driven evaluation practice from 1990 to 2009. *American Journal of Evaluation*, *32*, 199–226.

Coryn, C.L.S., Schröter, D. C., & Hanssen, C. E. (2009). Adding a time-series design element to the Success Case Method to improve methodological rigor: An application for nonprofit program evaluation. *American Journal of Evaluation*, *30*, 80–92.

Eisner, E. W. (2004). The roots of connoisseurship and criticism: A personal journey. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 196–202). Thousand Oaks, CA: Sage.

Funnel, S. C., & Rogers, P. J. (2011). *Purposeful program theory: Making effective use of theories of changes and program logic models*. San Francisco, CA: Jossey-Bass.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–10.

Kaplan, R. S., & Norton, D. P. (1996). *The balanced scorecard: Translating strategy into action*. Boston, MA: Harvard Business School Press.

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications* (2nd ed.). Thousand Oaks, CA: Sage.

Rogers, P. J. (2000). Program theory: Not whether programs work but how they work. In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed., pp. 209–232). Norwell, MA: Kluwer.

Sanders, W. L., & Horn, S. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, *8*, 299–311.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Yates, B. T. (1996). *Analyzing costs, procedures, processes, and outcomes in human services*. Applied Social Research Methods Series, Vol. 42. Thousand Oaks, CA: Sage.

CHAPTER 7

IMPROVEMENT- AND ACCOUNTABILITY-ORIENTED EVALUATION APPROACHES

Improvement- and Accountability-Oriented Evaluation Defined

This chapter summarizes three approaches whose proponents stress the need to fully assess a program's value. The approaches are decision- and accountability-oriented studies, consumer-oriented studies, and accreditation and certification.

These approaches are expansive and ideally comprehensive in considering the full range of questions and criteria needed to assess a program in terms of merit, worth, impact, probity, importance, feasibility, cost, safety, equity, and other factors. These approaches often incorporate the assessed needs of a program's stakeholders as the foundational criteria for assessing a program's worth and, in general, are grounded in principles of democracy. In addition, evaluators employing these approaches usually reference all of the pertinent technical and economic criteria for judging the merit or quality of program plans and operations. Improvement- and accountability-oriented evaluations also look for all relevant outcomes, not just those keyed to program objectives. Thus, improvement- and accountability-oriented evaluations may have an enlightening orientation. Usually such evaluations are objectivist and assume an underlying reality in seeking definitive, unequivocal answers to evaluation questions. They use multiple qualitative and quantitative assessment methods to provide cross-checks on findings.

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **The definition and functions of improvement- and accountability-oriented evaluation approaches**
- **The general strengths and weaknesses of such evaluation approaches**
- **The identity and developers of three specific improvement- and accountability-oriented approaches: decision- and accountability-oriented studies, consumer-oriented studies, and accreditation and certification[1]**
- **Key characteristics, strengths, and weaknesses of each of the three approaches**
- **Key considerations in applying each of the three approaches**

Functions of Improvement- and Accountability-Oriented Approaches

The main functions of the three approaches, respectively, are to (1) foster improvement and accountability through informing and assessing program decisions; (2) assist consumers in making wise choices among optional programs, products, and services; and (3) help accrediting associations certify meritorious institutions, programs, and personnel for service to clients, customers, beneficiaries, and others.

General Strengths and Weaknesses of Decision- and Accountability-Oriented Approaches

In general, the decision- and accountability-oriented approaches discussed in this chapter conform closely to this book's definition of evaluation because they are employed to fully assess a program's value. They involve addressing information needs for reporting credibly—to the full range of right-to-know audiences—on a program's soundness and accomplishments and also for issuing timely, focused evaluative feedback to help program staffs effectively carry out and strengthen programs. Proponents of the three approaches stress the importance of rigor and comprehensiveness in assessing programs and of employing an appropriate balance of qualitative and quantitative methods. Beyond their shared strengths, each of the three approaches also has unique strengths, as defined later in this chapter.

The main shared limitation of improvement- and accountability-oriented approaches is that overall they are oriented to providing all right-to-know audiences with a comprehensive, relatively long-term, definitive assessment of a program's merit and worth. This orientation calls for grounding evaluations in unambiguous and compelling definitions of a program's merit and worth, rigorous collection and analysis of the full range of relevant evidence, and issuance of pertinent reports to all right-to-know audiences. Some potential clients of evaluation can consider such approaches to be "overkill" in relation to what they see as their modest, short-term need for evaluative feedback and their possible unwillingness to share findings with a wide range of stakeholders. Also, the objective of fully assessing a program's merit and worth can be viewed as unrealistic because of the difficulty of achieving consensus on definitions of a program's merit and worth, gathering all relevant information on all pertinent variables, and weighting and combining the obtained pieces of information in a manner fully acceptable to the full range of stakeholders. Beyond these limitations of feasibility and the risk of overpromising in regard to what can be delivered, the three approaches have individual limitations and weaknesses as discussed later in the chapter.

Approach 15: Decision- and Accountability-Oriented Studies

The decision- and accountability-oriented approach is based on the premise that program evaluation should be used proactively to help improve a program as well as retrospectively to judge its value. The approach is distinguished from management information systems and from politically controlled studies mainly because decision- and accountability-oriented

studies emphasize questions of value. The approach's philosophical underpinnings include an objectivist orientation to finding best answers to context-limited questions and subscription to the principles of a well-functioning democratic society, especially human rights, an enlightened citizenry, equity, excellence, conservation, probity, and accountability. Practically, evaluators using this approach serve stakeholders by engaging them in focusing the evaluation and assessing draft evaluation reports and other materials; addressing their most important questions plus those required to assess the program's value; providing timely, relevant information to aid decision making and understanding; producing an accountability record; and issuing needed summative evaluation reports. This approach is best represented by the context, input, process, and product (CIPP) model for evaluation (Stufflebeam, 1967, 2003a, 2004b, 2005; also see Chapter 13 of this book), but elements of the approach are also seen in Cronbach's general approach to evaluation (1982). This section's discussion of the decision- and accountability-oriented approach is focused on the approach's widely used evaluation framework, the CIPP model.

Advance Organizers

Advance organizers of the decision- and accountability-oriented approach include decision makers and stakeholders; projected decisions to be made; program accountability requirements; and the criteria needed to examine a program's value—for example, its merit, worth, probity, feasibility, safety, importance, cost, and equity. Audiences include program decision makers and all other stakeholders, both internal and external to the program, such as recipients, business and institutional boards, parents and guardians, staff, administrators, program consultants, policymakers, funding authorities, and citizens. The decisions to be informed may include deciding to launch a program; determining the targeted recipients; defining goals and priorities; identifying and choosing from competing program strategies; planning procedures; scheduling, staffing, budgeting, and contracting the work; monitoring, adjusting, and reporting on operations; and deciding to continue, expand, contract, or terminate an effort.

Information for informing such decisions may be obtained by

- Assessing needs, problems, assets, opportunities, and objectives
- Identifying and assessing similar programs or alternative program approaches
- Assessing procedural plans, budgets, and schedules
- Assessing staff qualifications and performance
- Assessing program facilities and materials
- Monitoring and assessing program operations
- Assessing intended, unintended, short-range, and long-range outcomes
- Documenting and analyzing program costs
- Analyzing relationships between program resources, processes, and outcomes
- Comparing program outcomes and costs with those of similar programs

Purposes

The basic purpose of decision- and accountability-oriented studies is to provide a knowledge and value base for making and being accountable for decisions that result in developing, delivering, and making informed use of services that are both morally sound and cost effective. Evaluators must therefore interact with representative members of their audiences; discern their questions; determine appropriate criteria and information requirements (which may extend beyond the audiences' preferences); and report relevant, timely, efficient, and accurate information.

Under this approach, an evaluation's most important purpose is not to prove, but to improve. The improvement orientation means that evaluators seek to help a program mature, overcome its early deficiencies, and build on its strengths. However, improvement in a broader sense is sometimes best served by terminating a persistently ineffective program, thus freeing resources for better use. Although evaluators following this approach proactively foster and assist with ongoing improvement efforts, they also look retrospectively at what was attempted and accomplished. Thus, the approach is applied both formatively and summatively.

Stufflebeam's version (1967, 2003a, 2004b, 2005; also see Chapter 13 of this book) of this approach calls for evaluations to adhere to professional standards for evaluations. These include utility, feasibility, propriety, accuracy, and evaluation accountability (Joint Committee on Standards for Educational Evaluation, 2011).

Sources of Questions

The sources of questions addressed by the decision- and accountability-oriented approach's CIPP model are the concerned and involved stakeholders and the evaluator, the latter having a view of what questions must be addressed to assess a program's value. Stakeholders include all persons and groups involved in making choices or judgments related to initiating, planning, funding, staffing, implementing, and using a program.

A particular feature of the CIPP model is that it encourages and supports the notion of the evaluation client as a leader in the evaluation—a key program administrator who seeks, helps focus, and facilitates needed evaluation services, and who provides leadership throughout the course of a program to ensure the effective review and use of evaluation findings.

Questions

Illustrative questions for a formative evaluation are

- Has an appropriate beneficiary population been determined?
- What beneficiary needs should be addressed?
- What are the available alternative ways to address these needs, and what are their comparative benefits and costs?
- Are plans for services and participation morally defensible and technically sound?
- Are there adequate provisions in terms of facilities, materials, staff, and equipment?

- Are sufficient funds available to complete the program?
- Are program staff members sufficiently qualified and credible?
- Have appropriate roles been assigned to the different participants, and will they receive sufficient orientation and training?
- Are participants effectively carrying out their assignments?
- Is the program working well?
- What are the program's significant limitations and weaknesses?
- How, if at all, should the program be revised?
- What are the program's most important strengths?
- How might the program build on its strengths?
- Is the program effectively reaching all the targeted beneficiaries?
- Is the program meeting the participants' needs?
- Are recipients doing their part to make the program succeed?
- Is the program designed and functioning at least as well as its counterparts in other settings?

Primary questions for a summative evaluation are

- Did the program reach the targeted recipients and meet their pertinent needs?
- Is the program more cost effective than competing alternatives?
- What arrangements, events, and processes contributed to the program's success or failure?
- Did the program prove to be affordable?
- Is it beyond reproach?
- Is there a continued need for the program?
- Is it sustainable?
- Is it transportable?
- Was the program worth the investment?

The formative and summative questions are to be answered with respect to the underlying standards of good programs. Good programs must reach recipients and serve their targeted needs effectively, ethically, and at a reasonable cost, and they must perform as well as or better than reasonably available alternatives.

Methods

Many methods may be used in decision- and accountability-oriented program evaluations. These include, among others, document analysis, surveys, needs assessments, case studies,

competing advocacy/program design teams, carefully recorded and assessed observations, interviews, focus groups, resident evaluators, participant observers, cost analysis, and quasi-experimental and experimental designs.

To make the approach work, the evaluator must interact regularly with a representative group of stakeholders. In this respect, the approach is compatible with so-called participatory approaches to evaluation (Cousins & Earl, 1992; Cousins & Whitmore, 1998). Typically the evaluator at least should establish and engage a representative stakeholder review panel to help define evaluation questions, shape evaluation plans, facilitate information collection, review draft information collection instruments and reports, and help disseminate findings. The evaluator's exchanges with stakeholders involve conveying evaluative feedback that may be of use in program improvement, as well as determining what future evaluative feedback would be most helpful to program personnel and other stakeholders. Interim reports may assist beneficiaries, program staff, and others in assessing program operations and discerning problems requiring attention. By maintaining and accessing a dynamic baseline of evaluative information and applications of the information, the evaluator can periodically update the broad group of stakeholders on the program's progress, develop a comprehensive summative evaluation report, and supply program personnel with information for their own accountability reports.

The involvement of stakeholders is consistent with a key principle of the change process: an enterprise (in this case, an evaluation) can best effect change in a target group's behavior by involving members in planning, monitoring, and judging the enterprise. By involving stakeholders throughout the evaluation process, decision- and accountability-oriented evaluators lay the groundwork for helping stakeholders understand and value the evaluation process and apply the findings. Stakeholders' active participation in determining evaluation questions and procedures is also consistent with a principle of democracy, wherein citizens and stakeholders are given voice in decisions that will affect them.

Pioneers

Lee Cronbach (1963) advised educators to reorient their evaluations from a focus on objectives to a concern for making better program decisions. Although he did not use the terms *formative evaluation* and *summative evaluation,* he essentially identified and defined the underlying concepts before Scriven (1967) attached the now widely used labels to these concepts. In discussing the distinctions between the constructive, proactive orientation, on the one hand, and the retrospective, judgmental orientation, on the other, Cronbach (1963) argued for placing more emphasis on the former. He noted the limited functionality of the tradition of stressing retrospective outcome evaluation. In a later publication (Cronbach & Associates, 1980), Cronbach stressed that evaluations should take a long view and have an illuminating orientation. He saw evaluation's most important services to be enlightening societal groups about the workings of programs over time and informing policy development in key areas of societal need. (This conceptualization was a forerunner of the so-called realist evaluation approach [Henry, 2005], which calls for sustained, long-range study of how a particular

program strategy—such as Head Start—works out over decades in various national and international settings.)

Cronbach (1982) operationalized his evaluation approach in the UTOS model (also see Greene, 2004). In this model, the evaluator structures the evaluation to identify the units (U) targeted to receive program services, the program treatments (T) to be delivered, the observations (O) to be collected, and the settings (S) to be taken into account.

Following Cronbach's seminal call for evaluations to guide decision making, Stufflebeam (1966a, 1966b, 1967) also argued that evaluations should help program personnel make decisions keyed to meeting recipients' needs. Although he advocated an improvement orientation to evaluation, he also stressed that evaluators should both inform decisions and provide information for accountability (Stufflebeam, 1971a). He emphasized further that evaluators should interact with and report to the full range of stakeholders who need to make judgments and choices about a program. Stufflebeam's approach has been encapsulated in the CIPP evaluation model (Stufflebeam, 1967, 2003a, 2004b, 2005, 2007; Stufflebeam et al., 1971). That model (which is explained in detail in Chapter 13) calls for context, input, process, and product evaluations. Context evaluations involve assessment of pertinent needs, assets, opportunities, and problems to assist in formulating or judging goals and priorities. Input evaluations help identify and assess competing program strategies and procedural designs for meeting recipients' assessed needs. Process evaluations involve documenting and assessing the implementation of a selected program strategy. Product evaluations entail searching out, analyzing, and judging program results, in terms of such factors as reach to the targeted beneficiaries, effectiveness, side effects, sustainability, and transportability. (The CIPP Evaluation Model Checklist for implementing the CIPP model is available at www.wmich.edu/evalctr/checklists, and at www.josseybass.com/go/evalmodels, the Jossey-Bass Web site devoted to support of this book.)

Other contributors to the development of the decision- and accountability-orientated approach to evaluation are Marvin Alkin (1969) and William Webster (1975, 1995).

Use Considerations

The decision- and accountability-oriented approach is applicable in cases where program staff and other stakeholders require formative evaluation, summative evaluation, or both. It can provide the framework for both internal and external evaluations. When this approach is used for internal evaluation, it is often advisable to commission an independent metaevaluation of the inside evaluator's work (see Chapter 25). Beyond program evaluations, this approach has proved useful in evaluating personnel, students, projects, facilities, and products.

Contrary to some published misperceptions of the CIPP model's decision orientation, the model is not limited to serving the evaluation needs of high-level decision makers. On the contrary, evaluations following this model's precepts should, within reasonable bounds of feasibility, serve the evaluation-related information needs of the full range of a program's stakeholders. Clearly, stakeholders at all program levels have legitimate information needs related to the decisions they make, such as those concerning funding a program, directing

program operations, carrying out program procedures, or using the program's services. Virtually all program stakeholders also may legitimately expect an evaluation based on the CIPP model to provide them with information on a program's accountability—on whether or not the program involves prudent and effective use of funds, is meeting implementation milestones, is producing high-quality outcomes, and is delivering needed services to the targeted beneficiaries. Accordingly, evaluators who use this model should analyze, prioritize, and address stakeholders' decision- and accountability-oriented evaluative feedback needs at all levels of a program and throughout the program's duration.

One of the CIPP model's practical procedures for engaging and addressing the evaluation needs of all program stakeholders is the establishment of a broadly representative stakeholder review panel and engagement of the panel throughout the evaluation. This panel's functions include helping identify evaluation questions and information needs, reviewing draft evaluation tools and reports, facilitating the collection of needed evaluative information, and assisting with dissemination and application of evaluation findings.

Strengths

A major advantage of the decision- and accountability-oriented approach is that it encourages program personnel to use evaluation continuously and systematically to plan and implement programs that meet recipients' targeted needs. Its use aids decision making at all program levels, stresses improvement, and fosters quality assurance. It also presents a rationale and framework of information to help program personnel be accountable for their program decisions and actions. Its application involves the full range of stakeholders in the evaluation process to ensure that at least their highest-priority evaluation needs are addressed well and to encourage and support them in making effective use of evaluation findings. It is comprehensive in focusing on recipients' needs, program context, program plans and budgets, program operations, and program costs and outcomes. Although its provision for collecting needed information is comprehensive, the approach also is amenable to addressing a client's need for a study with a focused, narrow scope that might only assess a particular aspect of a program, such as its outcomes or the needs of targeted beneficiaries. For example, as appropriate, evaluators can pick and choose particular parts of the CIPP model that are most relevant to stakeholders' short-term needs. The decision- and accountability-oriented approach balances the use of quantitative and qualitative methods. It is keyed to professional standards for evaluations. Finally, proponents of the approach emphasize that evaluations must be grounded in the democratic principles of a free society and themselves be subject to evaluation.

Weaknesses

A weakness of the approach is that the collaboration required between an evaluator and stakeholders introduces opportunities for impeding the evaluation or biasing its results, especially when the evaluation situation is politically charged. Furthermore, when evaluators are actively influencing a program's course, they may identify so closely with the program that they lose

some of the independent, detached perspective needed to provide objective, forthright reports. Moreover, experience shows that the approach may overemphasize formative evaluation and give too little time and too few resources to long-term summative evaluation. Advance contracting, adherence to professional standards for evaluations, and external metaevaluation have been employed to counteract opportunities for bias and to ensure a proper balance of the formative and summative aspects of evaluation. Although the charge is erroneous, this approach carries the connotation that only top decision makers are served.

Approach 16: Consumer-Oriented Studies

In the consumer-oriented approach, the evaluator is the enlightened surrogate consumer (Scriven, 1994d). He or she must draw direct evaluative conclusions about the program being evaluated. Evaluation is viewed as the process of determining something's merit, worth, and significance, with evaluations being the products of that process (Scriven, 1991, 1993). The approach regards the consumer's welfare as a program's primary justification and affords that welfare the same primacy in program evaluation (see also Davidson, 2005). Grounded in a deeply reasoned view of ethics and the common good, and possessing skills in obtaining and synthesizing pertinent, valid, and reliable information, the evaluator should help developers produce and deliver products and services that are of excellent quality and of great use to consumers (for example, students, their parents, teachers, and taxpayers). More important, the evaluator should help consumers identify and assess the merit, worth, and significance of competing programs, services, and products. (The consumer-oriented approach is explained in detail in Chapter 14.)

Advance Organizers

Advance organizers include societal values, consumers' needs, costs, and criteria of goodness in the particular evaluation domain.

Purposes

The purpose of a consumer-oriented program evaluation is to judge the relative merit, worth, or significance of the products and services of alternative programs and thereby to help taxpayers, practitioners, and potential recipients make wise choices. The approach is objectivist in assuming an underlying reality and positing that it is possible, although often extremely difficult, to find best answers. It looks at a program comprehensively in terms of its quality and costs, functionally in regard to the assessed needs of the intended recipients, and comparatively considering reasonably available alternative programs. Evaluators are expected to subject their program evaluations to evaluation—that is, metaevaluation (Scriven, 1969b).

Sources of Questions

Questions for the consumer-oriented study are derived from society; program constituents (consumers); and especially the evaluator's frame of reference.

Questions

One general question is addressed: Which of several alternative programs is the best choice, given their differential costs, their levels of merit according to a range of criteria, the needs of the consumer group, the values of society at large, and evidence of both positive and negative outcomes?

Methods

Methods include checklists, needs assessments, goal-free evaluation, experimental and quasi-experimental designs, the modus operandi method, applying codes of ethical conduct, and cost analysis (Scriven, 1974). A preferred method is for an external, independent consumer advocate to conduct and report findings of a study of a publicly or privately supported program. The approach is keyed to employing a sound checklist of criteria pertaining to a program's main aspects. Scriven (1991) developed the generic Key Evaluation Checklist (KEC)—originally designed for evaluating educational products (Coryn, 2006)—for this purpose. (Regular updates of this checklist can be found at www.wmich.edu/evalctr/checklists.) The main evaluative acts in this approach are scoring, grading, ranking, apportioning, and producing the final synthesis (Scriven, 1994b, 2007).

The consumer-oriented approach employs a wide range of assessment topics. These include program description, background and context, consumers, resources, functions, delivery systems, values, standards, processes, outcomes, costs, critical competitors, generalizability, bottom-line assessment, practical significance, wide-ranging and long-term significance, recommendations, reports, and metaevaluation. The evaluation process begins with consideration of a broad range of such topics, supports continuous compilation of information on all of them, and ultimately culminates in a supercompressed judgment of the program's merit, worth, and/or significance. Largely, an evaluator using this approach applies Scriven's general logic of evaluation (1980), which consists of four steps, in arriving at a summative statement of merit or worth:

1. Establish criteria of merit: On what dimensions must the object of evaluation do well?
2. Construct standards: How well should the object of evaluation perform?
3. Measure performance and compare that with standards: How well did the object of evaluation perform?
4. Synthesize and integrate information into a judgment of merit, worth, or significance: What is the merit, worth, or significance of the object of the evaluation?

A consumer-oriented study requires a highly credible and competent expert, as well as either sufficient resources to allow the expert to conduct a thorough study or some other means of obtaining the needed information. Often a consumer-oriented evaluator is engaged to evaluate a program after its formative stages are over. In these situations, the external consumer-oriented evaluator is often dependent on being able to access a substantial base of information that program staff have accumulated. If no adequate base of information exists, the consumer-oriented evaluator will have great difficulty in obtaining enough information to produce a thorough, defensible summative program evaluation.

Pioneers

Michael Scriven (1967) was a pioneer in developing the consumer-oriented approach to program evaluation, and his work paralleled the concurrent work of Ralph Nader (1965) and Consumers Union in the general field of consumerism. Glass (1975) supported and developed Scriven's approach. Scriven (1967) coined the terms *formative evaluation* and *summative evaluation*. He noted that evaluations can be divergent in early quests for critical competitors and explorations related to clarifying goals and making programs function well. He also maintained, however, that evaluations ultimately must converge on summative judgments about a program's merit, worth, or significance. Although accepting the importance of formative evaluation, he also argued against Cronbach's position (1963) that formative evaluation should be given the major emphasis. According to Scriven (1991, 1993), the fundamental aim of a sound evaluation is to judge a program's merit, comparative value, wide-ranging significance, and overall worth. He sees evaluation as a transdiscipline encompassing all evaluations of various entities across all applied areas and disciplines and comprising a common logic, methodology, and theory that transcend specific evaluation domains, which also have their unique characteristics (Scriven, 1991, 1993, 2004a, 2004b; also see Coryn & Hattie, 2006).

Use Considerations

Given the emphasis on summative conclusions, consumer-oriented evaluations generally emphasize instrumental uses (that is, immediate decision making). Such evaluations are intended to inform a course of action for selecting among competing alternatives, considering the relative merit and/or worth of competing programs, services, or products. Less frequently, however, are such evaluations conducted with the intent of improving an evaluand. Unlike the decision- and accountability-oriented approach, consumer-oriented evaluations often do not directly involve relevant stakeholders and, therefore, decrease the likelihood of certain types of direct or indirect uses; they are not frequently conducted with the intent of improving existing programs, services, or products.

Strengths

One of the main advantages of a consumer-oriented evaluation is that it is a hard-hitting, independent assessment intended to protect consumers from shoddy programs, services, and products and to guide them to support and use those contributions that best and most cost-effectively address their needs. The approach's stress on independence and objectivity and its emphasis on achieving a comprehensive assessment of merit, worth, and significance have translated into high credibility with consumer groups. This is aided by the most up-to-date version of Scriven's KEC (available at www.wmich.edu/evalctr/checklists) and his *Evaluation Thesaurus* (1991), in which he presents and explains the checklist.

Weaknesses

One disadvantage of a consumer-oriented evaluation is that it can be so independent of program staff that it might not assist them in better serving consumers. Because stakeholders typically

are not meaningfully involved in planning a consumer-oriented evaluation, the evaluator is unlikely to identify, address, and differentially weight all the criteria that are of concern to different members of the evaluation's audience. Accordingly, some readers of the evaluation's final report may judge the evaluation as a failure in assessing the evaluand's most important characteristics and in producing useful findings. A summative evaluation that is conducted too early can intimidate developers and stifle their creativity. However, if such an approach is applied only near a program's end, the evaluator may have great difficulty in obtaining sufficient evidence to confidently and credibly judge a program's basic value. This often iconoclastic approach is also heavily dependent on a highly competent, independent, and "bulletproof" evaluator.

Approach 17: Accreditation and Certification

Many educational institutions, hospitals, and other service organizations have been the subject of an accreditation study, and many professionals, at one time or another, have had to meet certification requirements for a given position. Such studies of institutions and personnel are in the realm of accountability-oriented evaluations, and they also have an improvement element. Institutions, institutional programs, and personnel are studied (often using procedures similar to those used in auditing) to determine whether they meet requirements of given professions and service authorities and whether they are fit to serve designated functions in society (Chelimsky, 1985; Schwandt, 2005; Wisler, 1996). Typically the feedback reports identify areas for improvement.

Advance Organizers

The advance organizers used in an accreditation or certification study usually are guidelines and criteria that some accrediting or certifying body has adopted.

Purposes

The purpose of accreditation or certification studies is to determine whether institutions, institutional programs, or personnel should be approved to deliver specified public services.

Sources of Questions

The source of questions for accreditation or certification studies is the accrediting or certifying body.

Questions

Basically, accreditation or certification studies address these questions: (1) Are institutions and their programs or personnel meeting minimum standards? and (2) How can their performance be improved?

Methods

Typically an accreditation or certification study begins with self-study and self-reporting by the subject institution or individual. In the case of an institution, a panel of experts is assigned to visit the institution, verify a self-report, and gather additional information. Guidelines and criteria specified by the accrediting or certifying agency usually constitute the basis for the self-study and the visit by the expert panel. Typically, the accreditation process is concluded when the official accrediting board uses the self-study and visiting panel's reports to render a decision to fully accredit, accredit with conditions for improvement, or not accredit the subject program or organization. Typically accreditation is given for a limited period (such as five years).

Pioneers

Accreditation in education was pioneered by the College Entrance Examination Board around 1901. Since then, the accreditation function has been implemented and expanded, especially by the Cooperative Study of Secondary School Standards, dating from around 1933. Subsequently, the accreditation approach has been developed, further expanded, and administered by the North Central Association of Secondary Schools and Colleges, along with its associated regional accrediting organizations across the United States, and by many other accrediting and certifying bodies. Similar accreditation practices are found in medicine, law, architecture, and numerous other professions. Hughes and Kushner (2005) have provided a useful summary of the general approach to accreditation.

Use Considerations

Any area of professional service that potentially could put the public at risk—if services and products are not delivered by highly trained specialists in accordance with standards of good practice and safety—should consider subjecting its programs to accreditation reviews and its personnel to certification processes. Such use of evaluation services is very much in the public interest and is a means of getting feedback that can be used to strengthen capabilities and practices.

Strengths

The major advantage of accreditation or certification studies is that they aid consumers in making informed judgments about the quality of organizations and programs or the qualifications of individuals.

Weaknesses

Major difficulties with this approach are that the guidelines of accrediting and certifying bodies historically have often emphasized inputs and processes and given minimal attention to outcomes. However, over the past couple of decades accrediting organizations have given more attention to outcomes. Also, the self-study and visitation processes used in accreditation

offer many opportunities for corruption and inept assessment. Institutions have been known to present to evaluators only the program components they deem to be successful and to obscure program elements that are failing. Also, institutions have sometimes wined and dined visiting accreditation evaluators in the process of successfully co-opting them in the interest of getting favorable reports. As is the case for all other evaluation approaches, accreditation and certification processes should be subjected to independent metaevaluations keyed to the standards of sound evaluations. Unfortunately, individual accreditation processes are rarely subjected to independent metaevaluations.

Summary

In this chapter we have done the following:

- Identified three improvement- and accountability-oriented approaches: decision- and accountability-oriented studies, consumer-oriented studies, and accreditation and certification
- Identified, summarized, analyzed, and assessed Stufflebeam's CIPP model, as the principal example of the decision- and accountability-oriented approach
- Identified, summarized, analyzed, and assessed Scriven's conceptualization of the consumer-oriented approach, as the principal example from the category of consumer-oriented studies
- Discussed the general approach followed in accreditation or certification evaluations
- Noted that the three improvement- and accountability-oriented evaluation approaches emphasize the assessment of value, which is the thrust of the definition of evaluation used to classify the approaches considered in this book
- Stated that the three approaches all are aimed at serving both the public interest by assessing the soundness and value of programs and program developers' interests by providing feedback for effectively conducting and improving programs
- Noted that, in general, improvement- and accountability-oriented approaches are objectivist in their orientation
- Summarized each approach's unique characteristics and methods, as well as its strengths and weaknesses
- Referenced the availability for download of Stufflebeam's CIPP Evaluation Model Checklist and Scriven's Key Evaluation Checklist, and noted that these are useful tools for implementing the respective approaches
- Noted that all three approaches share the pervasive limitations and weaknesses inherent in seeking definitive conclusions about a program's value that may be unrealistic to obtain
- Reminded readers that Stufflebeam's decision- and accountability-oriented approach (the CIPP model) and Scriven's consumer-oriented approach are explained in detail in Chapters 13 and 14, respectively.

REVIEW QUESTIONS

1. Summarize this chapter's characterization of improvement- and accountability-oriented evaluation and state the essential differences between this type of evaluation and quasi-evaluation (defined in Chapter 6).
2. Based on this chapter, define the main thrust of the decision- and accountability-oriented evaluation approach—especially as embodied in the CIPP model—and list at least four main questions that this approach typically addresses.
3. Based on this chapter, what are the main functions of Scriven's consumer-oriented evaluation approach, and in what respects is his approach similar to and different from the approach that underlies *Consumer Reports* magazine?
4. A company has produced a catalogue of information technology equipment that promises to make any business office more efficient and cost effective and to improve customer relations. In the context of Scriven's consumer-oriented evaluation approach, what information would you require before placing an order from this catalogue?
5. Obtain a copy of Stufflebeam's CIPP Evaluation Model Checklist (available at www.wmich.edu/evalctr/checklists or www.josseybass.com/go/evalmodels), and apply it to plan an evaluation of a program with which you are familiar.
6. It is sometimes erroneously charged that the CIPP Evaluation model only or mainly addresses the evaluation needs of top decision-makers. Based on this chapter's characterization of the CIPP model, correct this misconception.
7. Explain, with examples, why Scriven has characterized the evaluation field as a transdiscipline.
8. Obtain a copy of Scriven's Key Evaluation Checklist (available at www.wmich.edu/evalctr/checklists), and apply it to plan an evaluation of a program with which you are familiar.
9. Explain this chapter's position that Stufflebeam's and Scriven's approaches have an objectivist orientation.
10. Explain and give illustrations of the essential services that accrediting bodies offer society.

Group Exercises

Work through the following exercises with your group. It is quite possible that members will reach different conclusions about the best responses to the presented assignments. However, members should try to reach a consensus or justify their opposing position. In advance of completing these exercises, it would be useful for group members to download and review Stufflebeam's CIPP Evaluation Model Checklist (www.wmich.edu/evalctr/checklists or www.josseybass.com/go/evalmodels) and Scriven's Key Evaluation Checklist (www.wmich.edu/evalctr/checklists).

Exercise 1

Based on this chapter, construct and fill in a four-by-six matrix showing how Stufflebeam's CIPP model and Scriven's consumer-oriented approach agree or disagree in regard to how they (1) define evaluation, (2) address the matter of involving stakeholders in the evaluation process, (3) accord relative emphasis to formative versus summative evaluation, (4) take into account the assessed needs of program beneficiaries, (5) employ quantitative as well as qualitative methods, and (6) define the range of program outcomes to be assessed. The four column headings for the assigned matrix are "Comparison Variables," "The CIPP Model," "Scriven's Consumer-Oriented Approach," and "Agree or Disagree" (to indicate where the two approaches come down on each of the six dimensions).

Exercise 2

Divide your group into Subgroup A and Subgroup B. Both subgroups should begin by reading the following paragraph.

A hospital's emergency room (ER) faces an impending accreditation review that is scheduled to commence about one year from now and is to be completed about eighteen months after that. The ER is required to complete an institutional self-study during the review's first twelve months. The accrediting organization will then conduct an external evaluation, including a review of the self-report, a subsequent site visit by a team of pertinent experts, and ultimately the accreditation agency's issuance of its summative evaluation report.

Subgroup A's Assignment

Reference the chapter's coverage of the CIPP evaluation model and review the CIPP Evaluation Model Checklist to draft a list of questions that the self-study should address.

Subgroup B's Assignment

Reference the chapter's coverage of Scriven's consumer-oriented approach and his KEC to draft a list of questions that the self-study should address.

Combined Group's Review and Deliberation

Each subgroup should present the other subgroup with its list of questions. After reviewing and contrasting the two lists of questions, the whole group should address the following questions:

- Which set of questions is better for guiding the ER's self-study, and why?
- Would a better set of questions be obtained by combining the two sets of questions? (If yes to this question, draft a synthesized list of questions.)
- To what extent was the CIPP Evaluation Model Checklist useful for determining salient questions?
- To what extent was the KEC useful for determining salient questions?
- What was the effect of not being able to interview stakeholders in the process of drafting the questions?

Exercise 3

Continuing with the assignment in exercise 2, as a whole group, list points in favor of building metaevaluation into the self-study. Likewise, list points in favor of subjecting the accreditation agency's external evaluation to an independent metaevaluation. Define the purposes of the metaevaluation of the self-study, and reach and defend a conclusion about whether this metaevaluation should be internal or external, or possibly both. Develop a rationale for a recommendation that the accrediting agency subject its external evaluation of the ER to an independent metaevaluation. (Please note that metaevaluation is defined in detail in Chapter 25, and that both the CIPP Evaluation Model Checklist and Scriven's KEC include sections on metaevaluation.)

Exercise 4

Cronbach and Scriven disagreed about the emphasis that should be given to formative evaluation and summative evaluation. Develop an evaluation scenario in which Cronbach's position makes more sense. Then develop an evaluation scenario that is more conducive to Scriven's position. Considering the two scenarios, write some guidelines to help evaluators decide when it is better to concentrate more on formative evaluation and when it is preferable to concentrate on summative evaluation.

Exercise 5

Briefly list the main strengths and weaknesses of the decision- and accountability-oriented approach to evaluation, especially as embodied in the CIPP model. Now state a situation in which you would find the approach highly useful, giving reasons. Then state another situation in which the approach either would not work or would not give as satisfactory an evaluation outcome as some other approach, again giving reasons. In regard to the latter situation, identify another approach that probably would work better than the decision- and accountability-oriented approach, again giving reasons.

Note

1. Stufflebeam's CIPP Model (a decision- and accountability-oriented approach) and Scriven's consumer-oriented approach are explained in detail in Chapters 13 and 14, respectively.

Suggested Supplemental Readings

Coryn, C.L.S. (2006). A conceptual framework for making evaluation support meaningful, useful, and valuable. *Evaluation Journal of Australasia*, *6*(1), 45–51.

Coryn, C.L.S., & Hattie, J. A. (2006). The transdisciplinary model of evaluation. *Journal of MultiDisciplinary Evaluation*, *3*(4), 107–114.

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, CA: Jossey-Bass.

Cronbach, L. J., & Associates. (1980). *Toward reform of program evaluation*. San Francisco, CA: Jossey-Bass.

Greene, J. C. (2004). The educative evaluator: An interpretation of Lee J. Cronbach's vision of evaluation. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 169–180). Thousand Oaks, CA: Sage.

Hughes, M., & Kushner, S. (2005). Accreditation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 4–7). Thousand Oaks, CA: Sage.

Joint Committee on Standards for Educational Evaluation. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Schwandt, T. A. (2005). Auditing. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 23–24). Thousand Oaks, CA: Sage.

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

Scriven, M. (1993). *Hard-won lessons in program evaluation*. New Directions for Program Evaluation, no. 58. San Francisco, CA: Jossey-Bass.

Scriven, M. (1994a). Evaluation as a discipline. *Studies in Educational Evaluation*, *20*, 147–166.

Scriven, M. (1994b). The final synthesis. *Evaluation Practice*, *15*, 367–382.

Scriven, M. (1994c). Product evaluation: The state of the art. *Evaluation Practice*, *15*, 45–62.

Scriven, M. (2004). Reflections. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 183–195). Thousand Oaks, CA: Sage.

Scriven, M. (2007). *Key Evaluation Checklist (KEC)*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/kec_feb07.pdf

Stufflebeam, D. L. (1971). The relevance of the CIPP evaluation model for educational accountability. *Journal of Research and Development in Education*, *5*(1), 19–25.

Stufflebeam, D. L. (2003). The CIPP model for evaluation. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 31–62). Norwell, MA: Kluwer.

Stufflebeam, D. L. (2004). The 21st century CIPP model: Origins, development, and use. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 245–266). Thousand Oaks, CA: Sage.

Stufflebeam, D. L. (2005). CIPP model (context, input, process, product). In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 60–65). Thousand Oaks, CA: Sage.

Stufflebeam, D. L. (2007). *CIPP Evaluation Model Checklist*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/cippchecklist_mar07.pdf

Wisler, C. (Ed.). (1996). *Evaluation and auditing: Prospects for convergence*. New Directions for Program Evaluation, no. 71. San Francisco, CA: Jossey-Bass.

CHAPTER 8

SOCIAL AGENDA AND ADVOCACY EVALUATION APPROACHES

Overview of Social Agenda and Advocacy Approaches

The social agenda and advocacy approaches are aimed at increasing social justice through program evaluation. Proponents of these approaches seek to ensure that all segments of society have equal access to educational and social opportunities and services. They advocate affirmative action to give the disadvantaged compensatory or preferential treatment through program evaluation. If, as many persons have stated, information is power, then these approaches to program evaluation are aimed at empowering the disenfranchised.

The four approaches in this set are oriented to employing the perspectives of stakeholders as well as those of experts in characterizing, investigating, and judging programs. They favor a constructivist orientation and the use of qualitative methods, with the exception of the constructivist and transformative evaluation approaches, which favor a mixed-method orientation. For the most part, they eschew the possibility of finding right or best answers and reflect the philosophy of postmodernism, with its attendant stress on cultural pluralism, moral relativity, and multiple realities. They provide for democratic engagement of stakeholders in obtaining and interpreting findings.

There is a concern that these approaches concentrate so heavily on serving a social mission that they might not meet the standards of sound evaluations. By giving stakeholders authority over key evaluation decisions, related

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **The central features of social agenda and advocacy approaches to evaluation**
- **The identity and developers of four social agenda and advocacy approaches: responsive or stakeholder-centered evaluation, constructivist evaluation, deliberative democratic evaluation, and transformative evaluation**
- **Key characteristics, strengths, and weaknesses of each of the four approaches**
- **Key considerations in applying each of the four approaches**

especially to the interpretation and release of findings, evaluators empower these persons to use evaluation to their advantage. Such delegation of authority over important evaluation matters can make the evaluation vulnerable to stakeholder bias and misuse of findings. Furthermore, if an evaluator is intent on serving the underprivileged, empowering the disenfranchised, or righting educational or social injustices, he or she might succumb to a conflict of interest and compromise the independent, impartial perspective needed to produce a dispassionate, valid assessment of a program's merit. For example, evaluators might be inclined to give a positive report on a corrupt program for the disadvantaged if funds allocated to serve these groups would be withdrawn as a consequence of a negative report. In the extreme, an advocacy evaluator could compromise the integrity of the evaluation process by working to achieve social objectives, and thus his or her study could devolve into a pseudoevaluation.

Nevertheless, there is much to recommend these approaches, because they are strongly oriented to democratic principles of equity and fairness and employ practical procedures for involving the full range of stakeholders. The particular social agenda and advocacy approaches presented in this chapter seem to have sufficient safeguards to walk the line between sound evaluation services and politically corrupt evaluations. Worries about bias control in these approaches underscore the importance of subjecting social agenda and advocacy evaluations, as well as all other types of evaluation, to independent metaevaluations grounded in standards for sound evaluations.

Approach 18: Responsive or Stakeholder-Centered Evaluation

The classic approach in this set is responsive evaluation, which is pluralistic, flexible, interactive, holistic, subjective, constructivist, and service oriented. The approach is relativistic because the evaluator seeks no final authoritative conclusion, interpreting findings against stakeholders' different and often conflicting values. The approach entails examining a program's full "countenance" (that is, the overall picture of the program as a whole) and prizes the collection and reporting of multiple diverse perspectives on the value of a program's orientation, operations, and achievements. Side effects and incidental gains as well as intended outcomes of a program are to be identified and examined.

The responsive approach has a strong philosophical base, whereby evaluators are expected to promote equity and fairness, help those with little power, thwart the misuse of power, expose hucksters, unnerve the assured, reassure the insecure, and always help people see things from alternative viewpoints. Advocates of the approach subscribe to moral relativity and posit that for any given set of findings, there are potentially numerous, conflicting interpretations that are equally plausible.

We refer to this approach using the term *stakeholder-centered evaluation* because one pervasive theme is that the evaluator must work with and for a diverse stakeholder group, including, for example, teachers, administrators, developers, taxpayers, legislators, and financial sponsors. The stakeholders are the "clients" in the sense that they support, develop, administer, or directly operate the program under study and seek or need the evaluator's counsel and advice in understanding, judging, improving, and using that program. The approach demands

that evaluators interact continuously with, and respond to, the evaluation needs of the various stakeholders, especially users of services. In doing so, the approach calls for inputs from experts as well as from the full range of program stakeholders.

Advance Organizers

The advance organizers in a responsive evaluation are stakeholders' concerns; issues in the program itself; as well as the program's rationale and background, and its intended and unintended transactions, outcomes, standards, and judgments.

Purposes

The responsive program evaluation may serve many purposes. Some of these are helping people in a local setting gain a perspective on a program's full countenance; understanding the ways that various groups see the program's problems, strengths, and weaknesses; and learning the ways affected people value the program, as well as the ways program experts judge it. The evaluator's process goal is to carry out a continuous search for key questions and standards and effectively communicate useful information to stakeholders as it becomes available. The responsive evaluation is intended to end with a perception of quality, not with action.

Sources of Questions

Key sources of questions for a responsive evaluation are both the full range of stakeholders and pertinent experts in the area of the evaluation. These may include community, practitioner, and beneficiary groups in the local environment, together with other interested parties outside the program's vicinity, plus specialists in the program area being assessed. The evaluator continually interacts with such persons to uncover and respond to their established and emergent concerns and the questions that they want addressed in the evaluation. In addition, the evaluator adds questions viewed as critically important and likely to be of interest to particular segments of the evaluation's intended audience.

Questions

In general, members of the audience usually want to know what the program has achieved, how it has operated, and how it has been judged by involved persons and experts in the program area. The more specific evaluation questions emerge as the study unfolds based on the evaluator's continuing interactions with stakeholders and program area experts and their collaborative assessment and interpretation of the developing evaluative information.

Typical questions that may emerge and be addressed in responsive evaluations include the following:

- What antecedents led to the subject program's development?
- What activities, interactions, allocations and expenditures of funds, and other transactions are significant and observable in the program's implementation?

- How do observed transactions compare with intended transactions?
- What are the program's outcomes—interim and long range, intended and unintended, positive and negative?
- What are the contingencies, including discernible causes and effects, between antecedents, transactions, and outcomes?
- How do various parties, including program stakeholders and relevant experts, judge the program?
- How can collected judgments be sorted in terms of absolute judgments (based only on one's conviction) and relative judgments (grounded in auditable evidence)?
- What is the range of value perspectives reflected in the various collected judgments of the program?

Methods

This approach reflects a formalization of the long-standing practice of informal, intuitive evaluation. It requires a relaxed and continuous exchange between evaluator and stakeholders. It is more divergent than convergent. Basically, the approach calls for continuing communication between evaluator and audience for the purposes of discovering, investigating, and addressing a program's issues.

Designs for responsive program evaluations are relatively open ended and emergent, building to narrative description rather than aggregating measurements across cases. The evaluator attempts to issue timely responses to stakeholders' concerns and questions by collecting and reporting useful information, even if the needed information was not anticipated at the study's beginning. Concomitant with the ongoing conversation with stakeholders, the evaluator attempts to obtain and present a rich set of information on the program. This includes its philosophical foundation and purposes, history, transactions, dilemmas, and outcomes. Special attention is given to side effects, the standards that various persons hold for the program, and their judgments of the program.

Depending on the evaluation's purpose, the evaluator may legitimately employ a range of methods. Preferred methods are case studies, expressive objectives, purposive sampling, observation, adversary reports, storytelling to convey complexity, sociodrama, and narrative reports. Responsive evaluators should check for the existence of stable and consistent findings by employing redundancy in their data collection activities and replicating their case studies (Stake, 1995). They are not expected to act as a program's sole or final judge, but they should collect, process, and report the opinions and judgments of the full range of stakeholders as well as those of pertinent experts. In the end, the evaluator makes a comprehensive statement of what the program is observed to be and references the satisfaction and dissatisfaction that appropriately selected people have in regard to the program. Overall, the responsive evaluator uses whatever information sources and techniques seem relevant for portraying the program's complexities and multiple realities and communicates the complexity even if the results instill doubt and make decision making more difficult.

Pioneers

Robert Stake (1967, 1975a, 1975b, 1983, 2003, 2004a, 2004b, 2011; Stake & Davis, 1999) is considered the pioneer of the responsive study, and his approach has been supported and developed by Denny (1978, 2011); Greene and Abma (2001); MacDonald (1975); Parlett and Hamilton (1972); Rippey (1973); and L. M. Smith and Pohland (1974). Guba's development of constructivist evaluation (1978) was heavily influenced by Stake's writings on responsive evaluation.

Use Considerations

The main condition for applying the responsive approach is a receptive client group and a confident, competent responsive evaluator. The client must be willing to endorse a quite open, flexible evaluation plan as opposed to a well-developed, detailed, preordinate plan; should expect budgetary requirements to unfold as the study develops; and should be receptive to equitable participation by a representative group of stakeholders. The client must find qualitative methods acceptable, and must usually be willing to forgo anything like a tightly controlled experimental study, although a controlled field experiment might be employed in exceptional cases. The client and other involved stakeholders need tolerance, even appreciation, for ambiguity, and should hold out only modest hope of obtaining definitive answers to evaluation questions. The clients must also be receptive to ambiguous findings, multiple interpretations, the existence of competing value perspectives, and the heavy involvement of stakeholders in interpreting and using findings. In this regard, the client should expect to assume responsibility for interpreting and applying findings. Finally, the client must be sufficiently patient to allow the program evaluation to unfold and find its direction based on ongoing interactions between the evaluator and the stakeholders.

This approach contrasts sharply with Scriven's consumer-oriented approach (1991, 2007), with its objectivist orientation and stress on reaching supercompressed summative judgments. Stake's evaluators are not the independent, objective assessors advocated by Scriven. The responsive or stakeholder-centered evaluator embraces local autonomy and helps people who are involved in a program to evaluate it and use the findings for program improvement. In a sense, the evaluator is a quite nondirective counselor who uses evaluation to help clients query about and gain insights into the workings of relevant projects and services and how these are addressing, or not addressing, targeted needs. Moreover, the responsive approach rejects the objectivist orientation, embodying instead the belief that there are no best answers and consistently preferable values as well as a preference for subjective information. In this approach, the program evaluation may culminate in conflicting findings and conclusions, leaving interpretation to the eyes of the beholders.

Strengths

A major strength of the responsive approach is that it involves action research in which people who are funding, implementing, and using programs are helped to conduct their own evaluations and use the findings to deliberate about concerns and issues in these programs

and improve their understanding, decisions, and actions. Responsive evaluators look deeply into stakeholders' main interests and search broadly for relevant information. In general, the evaluator and client study a program's mission and rationale, history, environment, transactions and operations, problems, and outcomes. They make effective use of qualitative methods and triangulate findings from different sources. The approach stresses the importance of searching widely for unintended as well as intended outcomes. It also gives weight to the meaningful participation in the evaluation by the full range of interested stakeholders, plus relevant experts. Judgments and other inputs from all such persons are respected and incorporated into the evaluation. The approach also provides for effective communication and assessment of findings through a range of techniques, such as focus groups, sociodramas, debates, and stories. Moreover, responsive evaluations help stakeholders develop limited, realistic expectations for what can be accomplished through systematic program evaluation.

Weaknesses

Stake (1967) has been modest, even pessimistic, in his claims about what evaluations can deliver. He has expressed skepticism about scientific inquiry as a dependable guide to developing generalizations about human services and pessimism about the potential benefits of formal program evaluations. He has expressed doubt that evaluators can contribute much to program improvement. But he has continued to write about evaluation and give advice on the subject because he recognizes that evaluations are going to be done and there is no option to close down the evaluation enterprise. In our informal exchange with Stake, he noted the following weaknesses: evaluation requires too much time for outside evaluators to get to know the program, it rejects the use of indicator variables and precision that many users want, it is too tolerant of subjectivity, it sacrifices reliability for too little gain in validity, and it is not conducive to recruiting and engaging psychometrists (whose assistance is needed). We think Stake might agree that his contributions are aimed at helping stakeholders make the best of what too often are unhelpful evaluations.

From our perspective, a major weakness of Stake's responsive approach is its vulnerability when it comes to external credibility, because people in the local setting have, in effect, considerable control over the evaluation of their own work. Similarly, evaluators working so closely with stakeholders may lose their independent perspective. As the evaluators advocate for those with little influence, those with authority and responsibility for the subject program may perceive the evaluators as lacking impartiality and evenhandedness. The approach is neither intended for nor amenable to reporting clear findings in time to meet decision or accountability deadlines. Moreover, rather than bringing closure, the approach's adversarial aspects and divergent qualities may generate confusion and contentious relations among stakeholders. Clients seeking definitive conclusions are unlikely to find them in a responsive evaluation report. Sometimes this evolving, exploratory approach may bog down an evaluation in an unproductive quest for multiple inputs and interpretations. Also, the divergent, open-ended nature of the approach makes for difficulties in budgeting and contracting the evaluation work.

Approach 19: Constructivist Evaluation

Egon Guba began developing the tenets of constructivist evaluation in the mid-1960s, and over the years published information on this approach under various labels, including *aexperimental design, naturalistic evaluation, effective evaluation,* and *fourth-generation evaluation*. Whatever the label, Guba grounded all renditions in a rejection of the principles and procedures of randomized controlled, variable-manipulating experimental design. Since the 1970s, when his wife, Yvonna Lincoln, joined him in developing this approach, they have regularly referred to the approach as fourth-generation evaluation (Guba & Lincoln, 1989). With this label they intended to convey the notion that their approach incorporates and goes beyond three earlier generations of evaluation approaches, which they characterized as focusing on objectives, description, and judgment, respectively. To these Guba and Lincoln added intensive participation of stakeholders in the design, conduct, reporting, and application of evaluations and also the constructions that different stakeholders bring to bear in judging a program. We see constructivism as the core concept in Guba and Lincoln's approach and thus are referring to their approach in this chapter as "constructivist evaluation."

The constructivist approach to program evaluation is heavily philosophical, service oriented, and paradigm driven. The constructivist evaluator rejects the tenets of logical positivism and instead embraces phenomenology and critical theory. Further, he or she rejects the existence of any ultimate reality and employs a subjectivist epistemology. Proponents of the approach see the knowledge to be gained as one or more social-psychological constructions that are uncertifiable, often multiple, and constantly problematic and changing. According to Lincoln (2005), the constructions are the "mental meanings, values, beliefs, and sense-making structures in which humans engage to make meaning from events, contexts, activities, and situations in their lives" (p. 162). Obtained constructions are to be treated holistically and analytically to reveal and study the underlying values, beliefs, and attitudes. Constructivist evaluation places the evaluator and program stakeholders at the center of the inquiry process, with all of them acting as the evaluation's "human instruments." Their focal activities are collecting, analyzing, and evaluating constructions.

Constructivist evaluation is as much recognizable for what it rejects as for what it proposes. In general, it strongly opposes positivism as a basis for evaluation, with its realist ontology, objectivist epistemology, and experimental method. It rejects any absolutist search for correct answers. It directly opposes the notion of value-free evaluation and attendant efforts to expunge human bias. It rejects positivism's deterministic and reductionist structure and its belief in the possibility of fully explaining studied programs. It also rejects requirements for impartiality that would preclude evaluators from advocating for stakeholders who are seriously disadvantaged and have little or no influence.

Advance Organizers

Advance organizers of the constructivist approach are basically the philosophical constraints placed on the study, as noted earlier, including the requirement of collaborative, unfolding

inquiry. A central advance organizer is the full range of program stakeholders, including especially those with few resources and little power and influence.

Purposes

The main purpose of constructivist evaluation is to determine and make sense of the variety of constructions that exist or emerge among stakeholders. Inquiry is kept open to ongoing communication and to the gathering, analysis, and synthesis of further constructions. One construction is not considered "truer" than others, but some may be judged as more informed and sophisticated than others. All evaluative conclusions are viewed as indeterminate, with the continuing possibility of finding better answers. All constructions are also context dependent. In this respect, the evaluator defines boundaries around what is being investigated.

Lincoln and Guba (Guba, 1978; Guba & Lincoln, 1981, 1989; Lincoln, 2003, 2005; Lincoln & Guba, 1985, 2004) proposed constructivist evaluation as a solution to problems they saw in evaluations based on classical experimental design. These problems include nonuse of findings, objectification of human beings, a lack of meaningful involvement of stakeholders in evaluations, and nonuse of evaluative processes by which people make sense of their world and the worlds of others.

Sources of Questions

The constructivist evaluator must respect participants' free will in all aspects of the inquiry process and should empower them to help shape and control the evaluation activities in their preferred ways. The evaluation must take account of the varying and often conflicting values of stakeholders. The approach requires explicit dialogue on values, particularly those in conflict. The inquiry process must also be consistent with effective ways of changing and improving society. Stakeholders must therefore play a key role in determining the evaluation questions, variables, and interpretive criteria. Evaluative foci include stakeholders' critical claims, concerns, and issues, as well as the program's objectives. Throughout the study, the evaluator regularly informs and consults stakeholders in all aspects of the evaluation. As findings emerge, they are shared widely with stakeholders. Constructivist evaluators need expertise in mediation, small- and large-group facilitation, and management.

The evaluator and stakeholders together identify the questions to be addressed. Constructivist evaluation insists that evaluators be totally ethical in respecting and advocating for all the participants in an evaluation, especially the disenfranchised. In shaping evaluation questions, evaluators are expected to help stakeholders take into account reasonably stable stakeholder characteristics, including gender, race, ethnicity, disability, socioeconomic status, cultural background, language, and sexual orientation. Values are held to be central in this evaluation approach, and strenuous measures are required both to take account of the full range of stakeholder values and to uncover relevant values that may not be apparent to stakeholders. Evaluators are authorized, even expected, to maneuver the evaluation to emancipate and empower involved or affected disenfranchised people in such spheres as civic engagement and democratic participation. Evaluators do this by raising stakeholders' consciousness so that they

are energized, informed, and assisted in transforming their world. Through epistemological exchanges, evaluators and stakeholders are expected to arrive at positions that are richly and deeply informed, factual, sophisticated, and nuanced.

Questions

The questions addressed in constructivist studies cannot be determined independently of participants' interactions. Questions emerge in the process of formulating and discussing the evaluation's purpose and the program's rationale, planning the schedule of discussions, and obtaining various stakeholders' initial views of the program to be evaluated. The questions develop further over the course of the approach's hermeneutic and dialectic processes. Questions may or may not cover the full range of issues involved in assessing something's merit and worth. The set of questions to be studied is never given in advance nor, after identification, considered fixed.

Methods

The constructivist methodology is first divergent, then convergent. Through the use of hermeneutics, the evaluator collects and describes alternative individual constructions pertaining to an evaluation question or issue, ensuring that each depiction meets with the respondent's approval. Communication channels are kept open throughout the inquiry process, and all respondents are encouraged and facilitated to make their inputs and are kept apprised of all aspects of the study. The evaluator then moves to a dialectic process aimed at achieving as much consensus as possible among different constructions. Respondents are provided with opportunities to review the full range of constructions along with other relevant information. The evaluator engages the respondents in a process of studying and contrasting existing constructions, considering relevant contextual and other information, reasoning out the differences among the constructions, and moving as far as they can toward a consensus. The constructivist evaluation is, in a sense, never ending. There is always more to learn, and finding ultimately correct answers is considered impossible.

Pioneers

As already noted, Guba and Lincoln (Guba, 1978; Guba & Lincoln, 1981, 1989; Lincoln, 2003, 2005; Lincoln & Guba, 1985, 2004) are pioneers in applying the constructivist approach to program evaluation. Harbans Bhola (1998), a disciple of Guba, has extensive experience in applying the constructivist approach to evaluating programs in Africa. In agreement with Guba, he has stressed that evaluations are always a function not only of the evaluator's approach and interactions with stakeholders but also of his or her personal history and outlook. Thomas A. Schwandt (1984, 1989), a student of Guba, has written extensively about the philosophical underpinnings of constructivist evaluation. Further, Fetterman's empowerment evaluation approach (2004, 2005) is closely aligned with constructivist evaluation, in that it seeks to engage and serve all stakeholders, especially those with little influence. However, there is a

key difference between the constructivist and empowerment evaluation approaches. Whereas the constructivist evaluator retains control of the evaluation and works with stakeholders to develop a consensus, the empowerment evaluator gives away authority over the evaluation to stakeholders while serving in a technical assistance role. This important distinction is a main reason why we classified empowerment evaluation as a type of pseudoevaluation.

Use Considerations

The constructivist approach can be applied usefully when evaluator, client, and stakeholders in a program fully agree that the approach is appropriate and pledge that they will cooperate. They should reach agreement on an understanding of what the approach can and cannot deliver. They need to accept that questions and issues to be studied will unfold throughout the process. They also should be willing to receive ambiguous and possibly contradictory findings, reflecting stakeholders' diverse perspectives. They should know that the shelf life of the findings is likely to be short (not unlike how it is with any other evaluation approach, but clearly acknowledged in the constructivist approach). They also need to value qualitative information that largely reflects the variety of stakeholders' perspectives and judgments. However, they should not expect to receive definitive pretest-posttest measures of outcomes or statistical conclusions about causes and effects. Although these persons can hope to achieve a consensus in the findings, they should agree that such a consensus might not emerge and that in any case, such a consensus would not necessarily generalize to other settings or time periods.

The approach rescinds any special privilege of scientific evaluators to work in secret and control or manipulate human subjects. In guiding the program evaluation, the evaluator balances verification with a quest for discovery, balances rigor with relevance, and balances the use of quantitative methods with the use of qualitative methods. The evaluator also prefers to provide rich and deep description rather than precise measurements and statistics. He or she employs a relativistic perspective to obtain and analyze findings, stressing locality and specificity over generalizability. The evaluator posits that ultimately there can be no correct conclusions. He or she exalts openness and the continuing search for more informed and illuminating constructions.

Guba and Lincoln (1989) have presented a set of criteria for judging constructivist evaluations that are analogous to scientific standards of rigor, validity, and value. The constructivist versions are credibility or trustworthiness, transferability beyond the studied context, dependability or reliability, and confirmability of data and data sources (see also Coryn, 2007b). One thrust of these criteria is that an evaluation's trustworthiness and utility are to be judged from the perspectives of the users of evaluation reports. Also, data are to be traced to their source and verified (for example, via member checks), and conclusions are to be assessed for logic, plausibility, and reasonableness.

In addition to these fairly standard criteria of sound inquiry, Lincoln and Guba (1985) have presented others that are intrinsic to constructivist evaluation. Called "authenticity criteria," they are balance and fairness in the evaluation report (Do evaluation reports present program strengths as well as weaknesses and fairly represent the views of all stakeholders?), ontological

authenticity (Did the evaluation help stakeholders understand their unconscious or unstated beliefs and values?), educative authenticity (Did the evaluation help stakeholders understand each other's perspectives and value positions?), catalytic authenticity (Did the evaluation prompt stakeholders to take actions?), and tactical authenticity (Did the evaluator effectively advocate for all stakeholders, including especially those with low levels of skill and influence?).

Strengths

This approach has a number of advantages. It is exemplary in fully disclosing the whole evaluation process and its findings. It is consistent with the principle of effective change that people are more likely to value and use an evaluation or any other change process if they are consulted and involved in its development: an evaluator using this approach seeks to directly involve the full range of stakeholders who might be harmed or helped by the evaluation as important, empowered partners in the evaluation enterprise. It is said to be educative for all the participants, whether or not a consensus is reached. It also lowers expectations for what clients can learn about causes and effects. Although it does not promise final answers, it moves from a divergent stage, in which there is a wide search for insights and judgments, to a convergent stage, in which some unified answers are sought. In addition, it uses participants as instruments in the evaluation, thus taking advantage of their relevant experiences, knowledge, and value perspectives, thus greatly reducing the burden of developing, field-testing, and validating information collection instruments before using them. The approach involves the effective use of qualitative methods and triangulation of findings from different sources.

Weaknesses

The approach is, however, limited in its applicability and has some disadvantages. Its openness and exploratory and participatory nature make it difficult to plan and budget for the required extensive and time-consuming evaluation process. Because of the need for full stakeholder involvement and ongoing interaction through both the divergent and convergent stages, it is often difficult to produce the timely reports that funding organizations and decision makers demand. Furthermore, if the approach is to work well, it requires the attention and responsible, continued participation of a wide range of stakeholders. The approach seems to be unrealistically utopian in this regard: widespread, grassroots interest and participation are often hard to obtain and especially to sustain throughout a program evaluation. Although the process emphasizes and promises openness and full disclosure, some participants do not want to tell their private thoughts and judgments to the world. Moreover, stakeholders sometimes are poorly informed about the issues being addressed in an evaluation and thus are poor data sources. All stakeholders are considered key data collection instruments in constructivist evaluations, but it is impractical to calibrate them to ensure they will carefully formulate and report valid observations and judgments. It can be unrealistic to expect that the evaluator can and will take the needed time to inform and then meaningfully involve those who begin basically ignorant of the program being assessed. Furthermore, constructivist evaluations can be greatly burdened by itinerant evaluation stakeholders

who come and go, reopen questions previously addressed, and question any consensus previously reached.

There is the further issue that some evaluation clients do not take kindly to evaluators who are likely to report competing, perspectivist answers and not take a stand concerning a program's merit and worth. Many clients are not attuned to the constructivist philosophy, and they may value reports that mainly include hard data on outcomes, assessments of statistical significance, and calibrated judgments. They may expect reports to be based on relatively independent perspectives that are free of program participants' conflicts of interest. Because the constructivist approach is a countermeasure to assigning responsibility for successes and failures in a program to certain individuals or groups, many policy boards, administrators, and financial sponsors might see this as an unworkable and unacceptable rejection of accountability. Finally, it is easy to say that all persons in a program should share the glory or the disgrace; but try to tell this to an exceptionally hardworking and effective teacher in a school program in which virtually no one else tries or succeeds.

Approach 20: Deliberative Democratic Evaluation

One of the most recent entries in the program evaluation models enterprise is the deliberative democratic evaluation approach advanced by House and Howe (2000a, 2000c, 2003) and House (2004, 2005). The approach functions within an explicit democratic framework and expects evaluators to uphold democratic principles in reaching defensible conclusions (Ryan, 2004, 2005). In this approach, program evaluation is envisioned as a principled, influential societal institution, contributing to democratization through the issuance of reliable and valid claims.

Advance Organizers

The advance organizers of deliberative democratic evaluation are seen in its three main dimensions: democratic participation, dialogue to examine and authenticate stakeholders' inputs, and deliberation to arrive at a defensible assessment of a program's merit and worth. House and Howe (2000a, 2000c, 2003) have deemed all three dimensions essential in all aspects of a sound program evaluation.

In the democratic dimension, the evaluator proactively identifies and arranges for the equitable participation of all interested stakeholders throughout the course of the evaluation. Equity is stressed, and power imbalances in which the views of powerful parties would dominate the evaluation message are not tolerated. In the dialogic dimension, the evaluator engages stakeholders and other audiences to assist in compiling preliminary findings. Subsequently, the collaborators seriously discuss and debate the draft findings to ensure that no participant's views are misrepresented. In the culminating deliberative dimension, the evaluator honestly considers and discusses with others all inputs obtained, but then renders what he or she considers a fully defensible assessment of the program's merit and worth. All interested stakeholders are given voice in the evaluation, and the evaluator acknowledges their views in the final report, but he or she may, in exercising professional discretion, express disagreement with some of them.

The deliberative dimension sees the evaluator reaching a reasoned conclusion by reviewing all inputs; debating them with stakeholders and others; reflecting deeply on all the inputs; and then reaching a defensible, well-justified conclusion.

Purposes

The purpose of the approach is to employ democratic participation in the process of arriving at a defensible assessment of a program.

Sources of Questions

The evaluator determines the evaluation questions to be addressed, but does so through dialogue and deliberation with engaged stakeholders.

Questions

Presumably, the bottom-line questions concern judgments about the program's merit and its worth to stakeholders.

Methods

Methods employed may include discussions with stakeholders, surveys, and debates. Inclusion, dialogue, and deliberation are considered relevant at all stages of an evaluation: inception, design, implementation, analysis, synthesis, write-up, presentation, and discussion. House and Howe (1998) presented the following ten questions for assessing the adequacy of a deliberative democratic evaluation:

- Whose interests are represented?
- Are major stakeholders represented?
- Are any excluded?
- Are there serious power imbalances?
- Are there procedures to control imbalances?
- How do people participate in the evaluation?
- How authentic is their participation?
- How involved is their interaction?
- Is there reflective deliberation?
- How considered and extended is the deliberation?

Pioneers

Ernest House is the originator of this approach. He and Kenneth Howe have said that many evaluators already implement their proposed principles, and have pointed to a monograph by

Karlsson (1998) to illustrate their approach. They also have referred to a number of authors who have proposed practices that at least in part are compatible with the deliberative democratic approach.

Use Considerations

The approach is applicable when a client agrees to fund an evaluation that requires democratic participation of at least a representative group of stakeholders. The funding agent must therefore be willing to give up sufficient power to allow inputs from a wide range of stakeholders, early disclosure of preliminary findings to all interested parties, and opportunities for the stakeholders to play an influential role in reaching the final conclusions. A representative group of stakeholders must be willing to engage in open and meaningful dialogue and deliberation at all stages of the study.

Strengths

The approach has many advantages. It is a direct attempt to make evaluations just. It involves the pursuit of democratic participation of stakeholders at all stages of the evaluation, ideally incorporating the views of all interested parties, including insiders and outsiders, disenfranchised persons and groups, as well as those who control the purse strings. Meaningful democratic involvement should direct the evaluation to the issues that people care about and make stakeholders inclined to respect and use the evaluation findings. The deliberative democratic evaluator employs dialogue to examine and authenticate stakeholders' inputs. A key advantage over some other advocacy approaches is that in this case the evaluator expressly reserves the right to rule out inputs that are considered incorrect or unethical. The evaluator is open to the views of all stakeholders, carefully considers them, but then renders a professional judgment of the program that is as defensible 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 or misinformed; nor does he or she necessarily leave that responsibility to the client. In rendering a final judgment, the evaluator ensures closure and an independent arrival at his or her own conclusions.

Weaknesses

As House and Howe (1998, 2003) have acknowledged, the deliberative democratic approach, pending further development and testing, is unrealistic and often cannot be fully applied. The approach—in offering and expecting full democratic participation to make an evaluation work—reminds us of a colleague who used to despair of ever changing or improving higher education. He would say that changing any aspect of a university would require getting every professor to withhold her or his veto. In view of the ambitious demands of the deliberative democratic approach, House and Howe have proposed it as an ideal to be kept in mind, although evaluators will seldom, if ever, be able to achieve it.

Approach 21: Transformative Evaluation

The final entry in the social agenda and advocacy evaluation approaches is transformative evaluation, which largely has been developed by Donna Mertens (1999, 2001, 2003, 2005b, 2007a, 2009; Mertens, Farley, Singleton, & Madison, 1994; Ryan, Greene, Lincoln, Mathison, & Mertens, 1998).

Advance Organizers

Although there are no clearly delineated advance organizers, broadly, the transformative paradigm is a philosophical framework for addressing issues of social justice, such as distributive justice and equity theory (see T. R. Tyler, Boeckmann, Smith, & Huo, 1997), through research and evaluation. It is grounded in multiple ideologies (feminism, participatory action research, resilience theory, positive psychology, critical race theory, and others) and is premised on the proposition that all knowledge claims are situational. One of the major principles underlying transformative evaluation is belief in the strength that often is overlooked in the grass roots of communities attempting to rectify intractable social problems.

Purposes

Historically, the transformative paradigm arose, in part, to address the growing dissatisfaction of feminists and critical theorists with the reality that the majority of sociological and psychological theory was developed by white males and based on studies of males (Gilligan, 1982). Given that the transformative evaluation approach recognizes the situational nature of knowing and knowledge claims, it would appear, on the surface, that it shares certain characteristics with constructivist evaluation, including the purpose of social justice.

Mertens (2005a), however, made a clear distinction between the two:

> The constructivist paradigm has been criticized not only by positivists and post-positivists, but also by another group of researchers representing a third paradigm of research: the transformative paradigm . . . .Transformative researchers argue that the constructivist paradigm did change the rules; however, it did not change the nature of the game. Constructivist researchers still consist of a relatively small group of powerful experts doing work on a larger number of relatively powerless research subjects. The transformative paradigm directly addresses the politics in research by confronting social oppression . . . [T]ransformative researchers go beyond the issue of the powerful sharing power with the powerless and relinquish control to the marginalized groups. (pp. 16–17)

Despite these differences, transformative evaluation is similar to constructivist evaluation in that both recognize the existence of multiple realities and ways of knowing. But unlike constructivist evaluation, the transformative evaluation approach places a special emphasis on the influence of societal, political, cultural, economic, ethnic, gender-related, and

disability-related values in constructing reality (Tarsilla, 2010a). "In addition, it emphasizes that that which seems 'real' may instead be reified structures that are taken to be real because of historical situations" (Mertens, 2005a, p. 23).

Sources of Questions

In transformative evaluation, the relationship between the evaluator and program participants is interactive. Further, the relationship should be empowering to those without power, and the evaluator should consider ways in which the evaluation benefits, or does not benefit, participants. Under this approach, it is an ethical imperative that inequalities be addressed by giving precedence to the voice of the least advantaged groups in society (Mertens, 2007a). Such groups thus are the key sources of questions for transformative evaluations.

Questions

Under this approach there are no illustrative, typical questions that can be identified in advance. The questions addressed by a transformative evaluation are determined by the marginalized groups who need to be served by the subject program. Such determinations require intense, sustained interactions between the evaluator and representatives of all segments of the overall group of stakeholders.

Methods

The transformative approach is methodologically pluralistic, and Mertens and many others (2005a, 2007b; see also Creswell & Plano Clark, 2007; Greene, 2007; Tashakkori & Teddlie, 2003) generally have advocated the use of mixed methods of inquiry, drawing from the repertoire of both the quantitative and qualitative traditions, in executing such evaluations.

Pioneers

Mertens (1999, 2001, 2003, 2005b, 2007a, 2009) is clearly the leader of the transformative movement in evaluation. Although she has led the development of the transformative evaluation approach, however, many, both inside and outside of evaluation, have influenced her views (see Mathison, 2005b).

Use Considerations

In designing and implementing a transformative evaluation, it is viewed as essential to include participants in all stages of the evaluation: planning, conduct, analysis, interpretation, and use of findings. Participants in such evaluations are not merely sources of information, but rather voices to be included in the entire evaluation process. In addition, transformative evaluation requires that evaluators be culturally competent (Mertens, 2005a; see also Evergreen & Cullen, 2010; Evergreen & Robertson, 2010; Tarsilla, 2010b; Thompson-Robinson, Hopson, SenGupta, 2004).

Strengths

The social justice emphasis in transformative evaluation is laudable, as is the call for evaluators to develop and apply cultural competence. Mertens (personal communication, April 12, 2011) pointed to Gunter and Rayner (2007); Habashi and Worley (2009); Hodgkin (2008); and Mertens, Harris, Holmes, and Brandt (2007) as exemplars of transformative research and evaluation in practice.

Weaknesses

As with the constructivist evaluation approach, transformative evaluation is likely to have limited credibility with some stakeholder groups. Whereas the disenfranchised groups probably would endorse the approach and its fairness to them, other important stakeholder groups, such as funders and the public at large, might view such studies as being too much under the control of particular stakeholder interest groups rather than professional evaluators. It may be hard to convince report recipients that an evaluation was sufficiently rigorous and unbiased for its findings to be taken seriously and used. Also, those responsible for program funding and oversight might question whether the transformative evaluation has fully, accurately, and effectively addressed typical evaluative questions concerned with such matters as assessed needs of targeted beneficiaries, documentation and assessment of program execution, program costs, and identification and assessment of program side effects as well as main effects.

Clearly, the transformative evaluation approach has addressed an important gap in the theory and practice of program evaluation. But like all other approaches, it carries significant limitations.

Summary

The four social agenda and advocacy evaluation approaches reviewed in this chapter are responsive or stakeholder-centered evaluation, constructivist evaluation, deliberative democratic evaluation, and transformative evaluation. They share the seeking of social justice through evaluation. All four approaches require extensive, sustained efforts to engage stakeholders in the evaluation process; they tend to favor the use of qualitative methods; and, with the exception of deliberative democratic evaluation, they embrace relativism and reject an objectivist perspective. Their shared strengths are their stress on engaging all stakeholders and their strong orientation to social justice. Their shared weaknesses have to do with problems of feasibility—for example, trouble sustaining full participation of all stakeholders throughout an evaluation process, difficulties in contracting for a loosely projected evaluation process, and many clients' insistence on obtaining bottom-line conclusions about a program's merit and worth. Although the four approaches are not exhaustive of social agenda and advocacy approaches, we think they represent such approaches' main themes. Overall, they provide valuable direction for evaluators who seek to meaningfully engage stakeholders and to pursue social justice through evaluations, while maintaining integrity in the evaluation work.

Space limitations precluded our inclusion of such other social agenda and advocacy entries as appreciative inquiry (Cooperrider & Whitney, 2005; Grant & Humphries, 2006; Preskill, 2005; Preskill & Catsambas, 2006; Preskill & Coghlan, 2003); critical theory evaluation (Freeman, 2010; MacNeil, 2005); feminist evaluation (Seigart, 2005; Seigart & Brisolara, 2002); illuminative evaluation (Hamilton, 2005); and approaches concerned with lesbian, gay, bisexual, and transgender issues (Cassaro, 2005). We also excluded consideration of the empowerment evaluation approach (Fetterman, 2005). Although empowerment evaluation clearly is a social agenda and advocacy approach, as explained in Chapter 5, we think it crosses the line into the pseudoevaluation category.

REVIEW QUESTIONS

1. What is the core mission of social agenda and advocacy approaches to evaluation?
2. What is meant by the claim that social agenda and advocacy approaches to evaluation have an affirmative action orientation?
3. What are two particular threats to the validity of social agenda and advocacy evaluations, and what are the sources of these potential shortcomings?
4. What are two main virtues of social agenda and advocacy evaluation approaches?
5. List two particular questions—reflecting the potential weaknesses of deliberative democratic evaluation—that should receive special attention in conducting a metaevaluation of an application of this approach.
6. What are the similarities and differences between the four social agenda and advocacy approaches in regard to whether an evaluator should reach a bottom-line judgment of a program's value?
7. What reasons are given in this chapter for referring to responsive evaluation as "stakeholder centered"?
8. What are two sharp disagreements between Scriven's consumer-oriented evaluation approach and Stake's responsive or stakeholder-centered evaluation approach?
9. Constructivist evaluation rejects the principles of any aspect of experimental design. Assess the extent to which Guba and Lincoln have provided alternative procedures for ensuring rigor in a constructivist evaluation.
10. Why was the transformative approach created, and what are its unique features?

Group Exercises

Exercise 1

Suppose that your group has been approached by a city manager to evaluate a special parks and recreation program. Assume that the city manager knows about responsive evaluation and wants the evaluation to follow this approach. He has asked you to provide a cross-section of community

members, including members of the city council, with an orientation. The meeting's purposes would be to orient the interested community members to the tenets of responsive evaluation and solicit their support and participation. What main points would you present? In particular, how would you advance and defend the notion that the evaluation will be pluralistic and relativistic? Also, how would you define the roles of children, parents, program staff, city government officials, and other stakeholders in planning, conducting, and reporting and using findings from this evaluation? How would you explain the specific responsibilities that the various stakeholders will be expected to fulfill? How would you respond to a charge that the approach is heavily prone to bias and is unlikely to produce clear, trustworthy, actionable findings?

Exercise 2

Suppose, following your presentation in response to exercise 1, that the community's mayor steps in to reject, or at least ignores, what she has heard. On the advice of one of her staff members, she proposes that the evaluation employ House's deliberative democratic approach instead of Stake's responsive approach. She then boldly states her preferences for the projected evaluation as follows:

- The evaluators should deliver their report only to the city council.
- The reported conclusions should be grounded in objective information shown to be reliable and valid.
- The evaluation should assess the extent to which the subject program is more effective than similar programs.
- The report should present a clear conclusion on the program's success and its superiority to one or more similar programs.
- The criteria for program success should be the program's stated objectives.
- The report should contain clear recommendations for continuing or terminating the program.
- The overall evaluation should be based on a fixed-price contract and confined to delivering a single final report within six months.

From the perspective of supporting the city manager's desire for an evaluation with significant stakeholder engagement and buy-in, how would you respond to the mayor's hard line on requiring a very different type of evaluation that she assumes could be conducted according to the requirements of deliberative democratic evaluation?

Suggested Supplemental Readings

Bhola, H. S. (1998). Program evaluation for program renewal: A study of the National Literacy Program in Namibia (NLPN). *Studies in Educational Evaluation*, *24*, 303–330.

Cassaro, D. A. (2005). Lesbian, gay, bisexual, and transgender issues in evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 226–228). Thousand Oaks, CA: Sage.

Cooperrider, D. L., & Whitney, D. (2005). *Appreciative inquiry: A positive revolution in change*. San Francisco, CA: Berrett-Koehler.

Denny, T. (1978). *Storytelling and educational understanding* (Occasional Paper Series, Paper #12). Kalamazoo: Western Michigan University, Evaluation Center.

Denny, T. (2011). Storytelling and educational understanding. *Journal of MultiDisciplinary Evaluation*, *7*(15), 258–271.

Fetterman, D. M. (2004). Branching out or standing on a limb: Looking to our roots for insight. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 319–330). Thousand Oaks, CA: Sage.

Fetterman, D. M. (2005). Empowerment evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 125–129). Thousand Oaks, CA: Sage.

Freeman, M. (Ed.). (2010). *Critical social theory and evaluation practice*. New Directions for Evaluation, no. 127. San Francisco, CA: Jossey-Bass.

Greene, J. C., & Abma, T. A. (Eds.). (2001). *Responsive evaluation*. New Directions for Evaluation, no. 92. San Francisco, CA: Jossey-Bass.

Guba, E. G. (1978). *Toward a methodology of naturalistic inquiry in educational evaluation*. Los Angeles: University of California, Center for the Study of Evaluation.

Guba, E. G., & Lincoln, Y. S. (1981). *Effective evaluation*. San Francisco, CA: Jossey-Bass.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Thousand Oaks, CA: Sage.

Hamilton, D. (2005). Illuminative evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 191–194). Thousand Oaks, CA: Sage.

House, E. R. (2004). Intellectual history in evaluation. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 218–224). Thousand Oaks, CA: Sage.

House, E. R. (2005). Deliberative democratic evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 104–108). Thousand Oaks, CA: Sage.

House, E. R., & Howe, K. R. (1998). *Deliberative democratic evaluation in practice*. Boulder: University of Colorado.

House, E. R., & Howe, K. R. (2000a). Deliberative democratic evaluation. In K. E. Ryan & L. DeStefano (Eds.), *Evaluation as a democratic process: Promoting inclusion, dialogue, and deliberation* (pp. 3–12). New Directions for Evaluation, no. 85. San Francisco, CA: Jossey-Bass.

House, E. R., & Howe, K. R. (2000b). Deliberative democratic evaluation in practice. In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed., pp. 409–421). Norwell, MA: Kluwer.

House, E. R., & Howe, K. R. (2003). Deliberative democratic evaluation. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 79–100). Norwell, MA: Kluwer.

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* (pp. 21–38). New Directions for Evaluation, no. 77. San Francisco, CA: Jossey-Bass.

King, J. A. (2005). Participatory evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 291–294). Thousand Oaks, CA: Sage.

Lincoln, Y. S. (2003). Constructivist knowing, participatory ethics and responsive evaluation: A model for the 21st century. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 69–78). Norwell, MA: Kluwer.

Lincoln, Y. S. (2005). Fourth-generation evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 161–164). Thousand Oaks, CA: Sage.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Thousand Oaks, CA: Sage.

Lincoln, Y. S., & Guba, E. G. (2004). The roots of fourth generation evaluation. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 225–242). Thousand Oaks, CA: Sage.

MacDonald, B. (1975). Evaluation and the control of education. In D. Tawney (Ed.), *Evaluation: The state of the art* (pp. 125–136). London, UK: Schools Council.

MacNeil, C. (2005). Critical theory evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 92–94). Thousand Oaks, CA: Sage.

Mertens, D. M. (1999). Inclusive evaluation: Implications of transformative theory for evaluation. *American Journal of Evaluation*, *20*, 1–14.

Mertens, D. M. (2005). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods* (2nd ed.). Thousand Oaks, CA: Sage.

Mertens, D. M. (2005). Transformative paradigm. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 422–423). Thousand Oaks, CA: Sage.

Mertens, D. M. (2007). Transformative considerations: Inclusion and social justice. *American Journal of Evaluation*, *28*, 86–90.

Mertens, D. M. (2007). Transformative paradigm: Mixed methods and social justice. *Journal of Mixed Methods Research*, *1*, 212–225.

Mertens, D. M. (2009). *Transformative research and evaluation*. New York, NY: Guilford Press.

Parlett, M., & Hamilton, D. (1972). *Evaluation as illumination: A new approach to the study of innovatory programs*. Edinburgh, UK: University of Edinburgh, Centre for Research in the Educational Sciences.

Preskill, H. (2005). Appreciative inquiry. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 18–19). Thousand Oaks, CA: Sage.

Preskill, H., & Catsambas, T. T. (2006). *Reframing evaluation through appreciative inquiry*. Thousand Oaks, CA: Sage.

Preskill, H., & Coghlan, A. T. (Eds.). (2003). *Using appreciative inquiry in evaluation*. New Directions for Evaluation, no. 100. San Francisco, CA: Jossey-Bass.

Rippey, R. M. (Ed.). (1973). *Studies in transactional evaluation*. Berkeley, CA: McCutcheon.

Schwandt, T. A. (1984). *An examination of alternative models for socio-behavioral inquiry*. Unpublished doctoral dissertation, Indiana University, Bloomington.

Schwandt, T. A. (1989). Recapturing moral discourse in evaluation. *Educational Researcher*, *18*(8), 11–16.

Scriven, M. (2005). Empowerment evaluation principles in practice [Review of the book Empowerment evaluation principles in practice, by D. M. Fetterman & A. Wandersman (Eds.)]. *American Journal of Evaluation*, *26*, 415–417.

Seigart, D. (2005). Feminist evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 154–157). Thousand Oaks, CA: Sage.

Stake, R. E. (1967). The countenance of educational evaluation. *Teachers College Record*, *68*, 523–540.

Stake, R. E. (1975). *Program evaluation, particularly responsive evaluation* (Occasional Paper Series, Paper #5). Kalamazoo: Western Michigan University, Evaluation Center.

Stake, R. E. (1983). Program evaluation, particularly responsive evaluation. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and social services evaluation* (pp. 287–310). Norwell, MA: Kluwer.

Stake, R. E. (1995). *The art of case study research*. Thousand Oaks, CA: Sage.

Stake, R. E. (1999). Summary of evaluation of Reader Focused Writing for the Veteran's Benefits Administration. *American Journal of Evaluation*, *20*, 323–343.

Stake, R. E. (2003). Responsive evaluation. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 63–68). Norwell, MA: Kluwer.

Stake, R. E. (2004). Stake and responsive evaluation. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 203–217). Thousand Oaks, CA: Sage.

Stake, R. E. (2004). *Standards-based and responsive evaluation*. Thousand Oaks, CA: Sage.

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

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.

Tarsilla, M. (2010). Inclusiveness and social justice in evaluation: Can the transformative agenda really alter the status quo? A conversation with Donna M. Mertens. *Journal of MultiDisciplinary Evaluation*, *6*(14), 102–113.

CHAPTER 9

ECLECTIC EVALUATION APPROACHES

Overview of Eclectic Approaches

Some evaluation theorists have made no commitment to any particular evaluation philosophy, methodological approach, or social mission. Instead, they have advanced pragmatic approaches that involve drawing from and selectively applying ideas and procedures from a wide range of other evaluation approaches. Proponents of these eclectic evaluation approaches hold no allegiance to any recognized school of evaluation thought and its adherents, but select such doctrines as they wish from various schools and apply them to the study at hand. Eclectic evaluation theorists derive ideas, style, or taste from a broad and diverse range of sources. Their approaches are designed to accommodate the needs and preferences of a wide range of evaluation clients and evaluation assignments, often with the express aim of examining a program unconstrained by the parameters of a single model or approach. Accordingly, evaluators following eclectic approaches employ whatever philosophical base, conceptual framework, and procedures may be conducive to achieving particular evaluation objectives and fulfilling the desires of particular evaluation clients. Evaluators following an eclectic approach on different occasions may conduct a case study, a randomized experiment, a responsive evaluation, an objectives-based study, a decision-oriented evaluation, a connoisseurship study, or something else. More likely, they will selectively employ elements of several evaluation approaches for any given study. The eclectic evaluation approaches discussed in this chapter are distinguished from pseudoevaluations because the former are committed to satisfying criteria of technical soundness, whereas the latter are not.

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **The central features of eclectic approaches to evaluation**
- **The identity and developers of two eclectic approaches: utilization-focused evaluation and participatory evaluation**
- **Key characteristics, strengths, and weaknesses of each of the two approaches**
- **Key considerations in applying each of the two approaches**

Among the eclectic evaluation approaches discussed in the evaluation literature are Owen's evaluation forms approach (2004, 2006; Owen & Rogers, 1999); the cluster evaluation approach employed by the W. K. Kellogg Foundation (Millett, 1995; Russon, 2005; J. R. Sanders, 1997); Patton's utilization-focused evaluation approach (1997, 2003, 2004, 2005b, 2008); and various participatory forms of evaluation (Cousins, 2003; Cousins & Earl, 1992; Cousins & Whitmore, 1998; King, 2005; Whitmore, 1998).

In his evaluation forms approach, Owen (2004, 2006; Owen & Rogers, 1999) drew from the writings of Alkin (1985) on decision uses of evaluation; Scriven (1980) on the logic of evaluation and values clarification; Weiss (1983) and Guba and Lincoln (1989) on stakeholder involvement; Patton (1997) on the use of findings; Rossi and Freeman (1993) on tailoring evaluation methods; and Stufflebeam (1983) for an adaptation of the context, input, process, and product (CIPP) model to provide a general classification of evaluation types.

The W. K. Kellogg Foundation spawned its cluster evaluation approach to effect collaborative study of clusters of its funded projects. This approach involves meetings of project evaluators, application of group process techniques, and use of a variety of data collection and synthesis procedures. The aim of cluster evaluation is to facilitate collaboration across similar projects, promoting the selection and use of common procedures to identify outcomes across the projects. Clearly the foundation was seeking an efficient way to evaluate and learn from groups of similar projects within its large portfolio of projects.

The most highly developed and widely used of the eclectic evaluation approaches are Patton's utilization-focused evaluation (1997, 2008) and practical participatory evaluation as advocated by Cousins (2003; Cousins & Earl, 1992; Cousins & Whitmore, 1998). Evaluators employing either of these approaches draw from the full range of evaluation concepts and methods and use whatever is deemed relevant to secure meaningful use of findings in particular evaluations. In this chapter we offer extended discussions of utilization-focused evaluation and participatory evaluation as exemplars of eclectic evaluation approaches.

Approach 22: Utilization-Focused Evaluation

The utilization-focused approach is explicitly geared to ensuring that program evaluations make an impact (Patton, 1997, 2008). It is a process for making choices to guide an evaluation study in collaboration with a targeted group of priority users, selected from a broader set of stakeholders, in an effort to focus effectively on their intended uses of the evaluation. All aspects of a utilization-focused program evaluation are chosen and applied to help the targeted users obtain and make appropriate use of evaluation findings, and to maximize the likelihood that they will do so. Such studies are judged more for the difference they make in improving programs and influencing decisions and actions than for their elegance or technical excellence. Michael Patton, this approach's primary developer, has argued that no matter how good an evaluation report is, if it only sits on the shelf gathering dust, it will not contribute positively to program improvement and accountability.

In deciding where to place Patton's evaluation approach within the category system used in this book, it became clear that it does not fit exclusively in the quasi-evaluation

category, improvement- and accountability-oriented category, or social agenda and advocacy category. At first glance, the approach seems to fit quite well in the social agenda and advocacy category. It requires democratic participation of a targeted (but not necessarily representative) group of stakeholders whom it empowers to determine the evaluation questions and information needs. In this regard, the evaluator engages the audience to set the agenda for the evaluation to increase the likelihood that the findings will be used. However, utilization-focused evaluators do not necessarily advocate any particular social agenda, such as affirmative action to right injustices and better serve the poor. Although the approach is in agreement with the improvement- and accountability-oriented approaches in regard to guiding decisions, promoting impacts, and invoking the program evaluation standards of the Joint Committee on Standards for Educational Evaluation (1994, 2011), it does not quite fit there. It does not, for example, require assessments of merit and worth. In fact, Patton essentially has said that his approach is pragmatic and adaptable. In the interest of getting findings used, he will draw on any legitimate approach to evaluation, leaving out any parts that might impede the findings' intended use. For these reasons, we place utilization-focused evaluation in the eclectic category and see it as the prime example of such evaluation approaches.

Advance Organizers

The advance organizers of utilization-focused program evaluations are, in the abstract, the possible users and uses to be served. Working from this initial conception, the evaluator moves as directly as possible to identify in concrete terms the actual users to be served. Through careful and thorough analysis of stakeholders, the evaluator identifies the multiple and varied perspectives and interests that should be represented in the study. He or she then selects a group that is willing to pay the price of substantial involvement and represents the program's stakeholders. The evaluator then engages members of this client group to clarify why they need the evaluation; how they intend to apply its findings; how they think it should be conducted; and what types of reports (for example, oral, printed) should be provided. He or she facilitates users' choices by supplying a menu of possible uses, types of information, and forms of reports for the evaluation. This is done not to supply the choices, but to help the client group thoughtfully focus and shape the study.

Among the possible uses of evaluation findings contemplated in this approach are assessment of merit and worth, improvement, and generation of knowledge. Proponents of this approach also value the evaluation process itself, seeing it as helpful in enhancing shared understandings among stakeholders, bringing support to a program, promoting participation in it, and developing and strengthening organizational capacity. According to Patton (2008), when the evaluation process is sound and functional, a printed final report may not be needed.

Purposes

In deliberating with intended users, the evaluator emphasizes that the program evaluation's purpose must be to give them the information they need to fulfill their objectives. Such objectives may include socially valuable aims, such as combating problems of illiteracy, crime,

hunger, homelessness, unemployment, child abuse, spouse or partner abuse, substance abuse, illness, alienation, discrimination, malnourishment, pollution, or bureaucratic waste. However, it is the targeted users who determine the program to be evaluated, what information is required, how and when it must be reported, and how it will be used. Patton (2008) explicitly has not sold his approach as one aimed particularly at righting social wrongs, because he leaves evaluation objectives and outcomes to the client and users.

Sources of Questions

Orienting questions for a utilization-focused evaluation typically include, but are not limited to,

- What decisions, if any, are evaluation findings expected to influence?
- When will decisions be made, and by whom?
- When must evaluation findings be presented to be timely and influential?
- What is at stake in the decisions, and for whom?
- What controversies or issues surround the decisions?
- What are the history and context of the decision-making process?
- What other factors (values, politics, personalities, promises already made) will affect decision making?
- To what extent has the outcome of the decision already been determined?
- What data and findings are needed to support decision making?
- How will we know afterward if the evaluation was used as intended?

In this approach, the evaluator essentially serves as the intended users' technical assistant. Among other roles, he or she is a facilitator of stakeholders' decision making. The process of identifying and aiding relevant decision makers and those who will use information garnered from the evaluation is basic to utilization-focused evaluation, which is very much a participant-oriented approach. Patton (2008) has stated, however, that the evaluation should meet the full range of professional standards for program evaluations, not just the requirement of utility. It is hard for us to see how this aim is to be achieved, because the evaluator gives so much authority to users of the evaluation. His response is that the evaluator must be an effective negotiator, standing on principles of sound evaluation but working hard to gear a defensible program evaluation to the targeted users' evolving needs. The utilization-focused evaluation is considered situational and dynamic. Depending on the circumstances, the evaluator may play any of a variety of roles: trainer, planner, negotiator, facilitator, measurement expert, internal colleague, external expert, analyst, spokesperson, or mediator.

Questions

The evaluator works with the targeted users to determine the evaluation questions. Such questions are to be stipulated locally, may address any of a wide range of concerns, and

probably will change over time. Example foci are processes, outcomes, impacts, costs, and benefits. The chosen questions are kept at the forefront and provide the basis for guiding information collection and the reporting of plans and activities, so long as users continue to value and pay attention to the questions. Often, however, the evaluator and client group will adapt, change, or refine the questions as the evaluation unfolds.

Methods

All evaluation methods are fair game in a utilization-focused program evaluation. The evaluator will creatively employ whatever methods are relevant (quantitative and qualitative, formative and summative, naturalistic and experimental, and so on). As much as possible, the utilization-focused evaluator puts the client group in the driver's seat in determining evaluation methods to ensure that he or she focuses on their most important questions and employs methods they trust; collects the appropriate information; applies the relevant values; answers the key action-oriented questions; reports the information in such a form and at such a time as to maximize use; convinces stakeholders of the evaluation's integrity and accuracy; and facilitates the users' study, application, and—as appropriate—dissemination of findings. The bases for interpreting evaluation findings are the users' values, and the evaluator will engage in values clarification to ensure that evaluative information and interpretations serve users' purposes. Users are actively involved in interpreting findings. Throughout the evaluation process, the evaluator balances a concern for utility with provisions for validity and cost-effectiveness.

In general, the methodology of utilization-focused program evaluation is labeled "active-reactive-adaptive" and "situationally responsive," emphasizing that it evolves in response to ongoing deliberations between evaluator and client group and in consideration of contextual dynamics. Patton (1997) said,

> Evaluators are active in presenting to intended users their own best judgments about appropriate evaluation focus and methods; they are reactive in listening attentively and respectfully to others' concerns; and they are adaptive in finding ways to design evaluations that incorporate diverse interests . . . while meeting high standards of professional practice. (p. 383)

Pioneers

Patton (1980, 1982, 1994, 1997, 2004, 2005b, 2008) is the leading proponent of utilization-focused evaluation. Other advocates of the approach are Alkin (1995, 2011); Cronbach and Associates (1980); Davis and Salasin (1975); and the Joint Committee (1994, 2011).

Use Considerations

As defined by Patton, the approach has virtually universal applicability. It is situational and can be tailored to meet any program evaluation assignment. It carries with it the integrity of sound evaluation principles as defined in *The Program Evaluation Standards* (Joint Committee,

1994, 2011). Working within these general constraints, the evaluator negotiates all aspects of an evaluation to serve the specific individuals who need to have the program evaluation performed and who intend to make concrete use of the findings.

In utilization-focused evaluation the evaluator selects from the entire range of evaluation techniques those that best suit the particular evaluation. And he or she plays any of a wide range of evaluation and improvement-related roles that fit local needs. The approach requires a substantial outlay of time and resources by all participants, for conducting both the evaluation and the needed follow-through. Nonetheless, the methodological pluralism underlying the approach is aimed directly at reflecting the multiple realities that constitute programs.

Strengths

The approach is geared toward maximizing evaluation impacts. It fits well with a key principle of change: individuals are more likely to understand, value, and use the findings of an evaluation if they were meaningfully involved in the enterprise's planning and execution. As Patton (1997) said, "By actively involving primary intended users, the evaluator is training users in use, preparing the groundwork for use, and reinforcing the intended utility of the evaluation" (p. 22). The evaluator engages stakeholders to determine the evaluation's purpose and procedures and uses their involvement to promote the use of findings. It bears mention that evaluators using Patton's approach address stakeholder involvement more realistically than do evaluators employing some of the other evaluation approaches. Rather than trying to reach and work with all stakeholders, evaluators work with a selected group of users who agree to become and stay actively involved in the process. The approach emphasizes values clarification and paying close attention to contextual dynamics. It may selectively involve any and all relevant evaluation procedures, whether based on quantitative or qualitative methodology (or both), and findings from different sources may be triangulated. One significant value of the approach is that it may emphasize a formative role to foster and assist positive program development rather than a summative one, the latter of which can stifle exploration and creativity in a program's early stages. Finally, proponents of this sophisticated and socially acceptable approach stress the need to meet all relevant standards for evaluations.

Weaknesses

Patton (2008) sees the main limitation of the approach as the turnover of involved users. Bringing in replacement users may require that the program evaluation be renegotiated. This renegotiation may be necessary to sustain or renew the prospects for evaluation impacts, but it can also derail or greatly delay the process. Furthermore, it is easy to say that this approach should meet all of the program evaluation standards (Joint Committee, 1994, 2011), but it is hard to see how this can be accomplished with any consistency. The approach seems to be vulnerable to corruption by user groups, because they are given much control over what will be looked at, the questions addressed, the methods employed, and the information to be

collected. Moreover, it is often difficult to define and limit the user groups, the reasons for the evaluation, and the audiences for any reports. Stakeholders with personal, biased priorities and interests may unduly influence the evaluation, especially if they sustain active participation while other stakeholders with different but legitimate points of view and interests do not have time for sustained involvement. For example, a narrow interest group may limit the evaluation to a subset of questions that is too narrow. It may be almost impossible to get a representative group of users to agree to and follow through on a sufficient commitment of time and safeguards to ensure an ethical, valid process of data collection, reporting, and use. Moreover, effective implementation of this approach requires a highly competent, confident evaluator who can approach any situation flexibly without compromising basic professional standards. Strong skills of negotiation are essential, and the evaluator must possess expertise in the full range of quantitative and qualitative evaluation methods, strong communication and political skills, and working knowledge of all applicable standards for evaluations. Clearly, such an evaluator must be capable of conducting a "quick study" to promptly grasp a situation's programmatic and social dynamics and needs. Unfortunately, not many evaluators are sufficiently trained and experienced to meet these demanding requirements (Dewey, Montrosse, Schröter, Sullins, & Mattox, 2008; King, Stevahn, Ghere, & Minnema, 2001; Stevahn, King, Ghere, & Minnema, 2005).

Approach 23: Participatory Evaluation

Participatory forms of evaluation largely emerged from the action research and rapid rural appraisal paradigms, among others, and were largely formalized as an evaluation approach in the early 1980s (Brisolara, 1998; Brunner & Guzman, 1989; Chambers, 1992, 1994; Cullen, 2009). Philosophically, participatory evaluation's concern with participation is partly a response to "questions posed by the critique of orthodox social science practice that emerged during the 1960s and 1970s" (Brisolara, 1998, p. 27); the approach is an extension of the more restrictive stakeholder-centered evaluation (Bryk, 1983; Mark & Shotland, 1985).

Numerous terms have been used to describe participatory forms of evaluation, including *participatory rural appraisal, participatory action research, community-based participatory research,* and *asset-based community development* (Cullen & Coryn, 2011; Cullen, Coryn, & Rugh, 2011); *collaborative evaluation* and *inclusive evaluation* (Owen, 2004; Rodriguez-Campos, 2005; Ryan, Greene, Lincoln, Mathison, & Mertens, 1998); *empowerment evaluation* (Fetterman, 1994, 2001, 2004, 2005); *evaluation capacity building* (Preskill & Russ-Eft, 2005; Preskill & Torres, 1999a); and *practical participatory evaluation* and *transformative participatory evaluation* (Cousins & Whitmore, 1998), among many others. Often these terms are used interchangeably (Cullen, 2009; Cullen & Coryn, 2011; O'Sullivan & D'Agostino, 2002). As Cousins and Whitmore (1998) noted, "For some, [participatory evaluation] implies a practical approach to broadening participation in decision making and problem solving through systematic inquiry, for others, reallocating power in the production of knowledge and promoting

social change" (p. 5). According to Cousins (2003), there is little consensus on what is meant by participatory evaluation:

> Participatory evaluation (PE) turns out to be a variably used and ill-defined approach to evaluation that, juxtaposed to more conventional forms and approaches, has generated much controversy in educational and social and human services evaluation. Despite a relatively wide array of evaluation and evaluation-related activities subsumed by the term, evaluation scholars and practitioners continue to use it freely often with only passing mention of their own conception of it. There exists much confusion in the literature as to the meaning, nature, and form of PE and therefore the conditions under which it is most appropriate and the consequences to which it might be expected to lead. (p. 245)

Cousins, Donohue, and Bloom (1996) developed a widely cited framework for differentiating between different participatory forms of evaluation that was subsequently revised by Cousins and Whitmore (1998), and later by Weaver and Cousins (2004). According to the original framework, all participatory forms of evaluation can be classified along three dimensions: control of the evaluation process, stakeholder selection for participation (Which stakeholders are included in the evaluation?), and depth of participation (In what capacity do stakeholders participate?).

Cullen et al. (2011) modified and extended the previously posited dimensions of the participatory evaluation process, and proposed four distinguishing features of the approach. The first, technical control, describes who is responsible for technical aspects of an evaluation (that is, the evaluator, stakeholders, or some combination of both). The second, extent of participation, describes the degree of participation by engaged stakeholders, ranging from consultation to deep participation. The third, stakeholder group, describes who participates (that is, selected users or all legitimate stakeholders). The final feature, phase of participation, describes whether or in what specific phases of an evaluation stakeholders participate (that is, during design, data collection, data analysis, interpretation, making recommendations, reporting, and/or dissemination). Daigneault and Jacob (2009) have developed methods for quantifying the degree of stakeholder participation in participatory forms of evaluation. Interested readers should also refer to Fitzpatrick, Sanders, and Worthen (2011) for additional details concerning the participatory evaluation approach.

Throughout this section, we mainly describe what Cousins and Whitmore (1998) have labeled "practical participatory evaluation," which predominately arose in the United States and Canada and which tends to incorporate principles of organizational learning and development (for example, Preskill, 1994). Cousins defined 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" (Cousins & Earl, 1992, p. 399). At its core, participatory evaluation is characterized as members of two different professional communities working in partnership) or a partnership between someone who is trained in evaluation methodology and those who are not.

Advance Organizers

The advance organizers of the participatory evaluation approach are neither particular methods nor particular questions, but rather the sometimes conflicting worldviews, experiences, perspectives, and values of multiple stakeholder groups with the specific intent of promoting buy-in, use of findings, and potential for change.

Notably, and again like Patton (1997, 2008) with his utilization-focused evaluation approach, Cousins (2003, 2004a) has endorsed sound evaluation principles as defined in *The Program Evaluation Standards* (Joint Committee, 1994, 2011) for participatory forms of evaluation. Accordingly, he has placed a strong emphasis on these standards, including in particular the requirement for stakeholder involvement during evaluation capacity–building activities and throughout the evaluation process. Cousins also has endorsed the standards' requirement of both internal and external formative and summative metaevaluations.

Purposes

Participatory forms of evaluation tend to emphasize program improvement rather than summative judgments of merit and/or worth. Participatory evaluators promote improvement by facilitating process use (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); conceptual use (whereby no direct action is taken based on an evaluation but understanding of a program has been affected); and symbolic use (whereby the mere existence of an evaluation, rather than any aspect of its results, is used to persuade or to convince), rather than instrumental use (involving instances in which evaluation knowledge is used directly). These forms of use, and others, are described in detail by Cousins (2004a, 2004b, 2007); Johnson et al. (2009); and Patton (1997, 2008).

Sources of Questions

Clearly, the evaluation's intended users are the main source of evaluation questions. Ultimately, the participatory evaluator works collaboratively in partnership with a selected group of intended users, and his or her tasks are to provide technical support and training and to ensure and maintain quality control. As with utilization-focused evaluation, a major emphasis of participatory evaluation is actively promoting the use of evaluation findings. However, the latter typically involves a broader group of stakeholders than the former.

Questions

The types of questions that may be addressed are wide ranging, restricted in their content only by the scope of the interests of the evaluation's stakeholders.

Methods

Similar to Patton's utilization-focused evaluation approach (1997, 2008), participatory forms of evaluation promote use through the involvement of selected stakeholders and intended users

throughout the entire evaluation process, based on the principle that including these groups will increase evaluation buy-in and ultimately the likelihood of evaluation use.

Also as with Patton's utilization-focused approach (1997, 2008), participatory evaluation is methodologically flexible and pluralistic rather than narrow and singular in focus. That is, the participatory evaluator draws from the entire range of available evaluation methods, which often are determined in collaboration with selected stakeholders.

In doing so, the participatory evaluator frequently begins by identifying a selected group of potential users and evaluation participants from a broad group of stakeholders. He or she then engages this group to clarify why they need the evaluation; what their most important questions or issues are; how they intend to apply findings; how they think the evaluation should be conducted; how findings will be disseminated; and how the evaluation's impact or program improvement will be defined, measured, and judged.

To address each of these points, the participatory evaluator engages in a variety of evaluation capacity–building activities (workshops and trainings, situational analyses, facilitated discussions between stakeholder groups, and so on). Then, or often simultaneously, the evaluator facilitates and guides the execution of the evaluation by providing selected stakeholders with technical assistance throughout the entire process. All the while, the evaluator makes a concerted effort to ensure technical quality and accuracy by guiding (but not necessarily prescribing) question formulation; evaluation design; the choice of methods; data-gathering strategies; appropriate analytic techniques and approaches; and clear, concise report writing.

Pioneers

J. Bradley Cousins is one of the most frequently cited, prolific, and influential theorists on participatory evaluation, and the contemporary participatory movement in evaluation is largely attributed to his work (Cousins, 1996, 2003, 2004b; Cousins & Earl, 1992, 1995; Cousins & Shulha, 2008; Cousins & Whitmore, 1998; Robinson & Cousins, 2004; Weaver & Cousins, 2004).

Strengths

Despite the numerous criticisms cited later, even many of the approach's detractors acknowledge its user-friendliness. Moreover, stakeholder participation is a part of nearly all evaluations. King (2007), for example, argued that "to a certain extent all program evaluation is participatory—evaluators must, after all, talk to someone when framing a study" (p. 83). It is, however, how stakeholders engage in participatory evaluation that distinguishes the participation inherent in the approach from more ordinary forms. Even so, in their research on participatory evaluation, Cousins et al. (1996) and Cullen et al. (2011) have found that even among those who self-identify as being participatory evaluators, the majority of their practices tend to align more closely with forms of stakeholder-centered evaluation in which stakeholder groups participate only to a limited degree. In addition, Cullen et al. found that evaluators working in international development settings who self-identified as being participatory evaluators frequently interpreted "interviewing stakeholders or gathering or eliciting other types of information (e.g., recipients, government officials, implementing partners) as a legitimate form

of participation" (p. 356). Cullen et al. argued that "this view treats the notion of participation as essentially using stakeholders as sources of information or data" (p. 356).

Although others (Garaway, 1995; B. Levin, 1993; Mark & Shotland, 1985) have provided various justifications for participatory forms of evaluation, Cousins et al. (1996) argued that the approach's main advantages are practical—it can be employed to obtain knowledge that is likely to be used, to secure stakeholders' buy-in, and to determine answers to questions that are relevant to selected stakeholders. Cousins (1996) and Greene (1988a, 1988b) have found that stakeholder participation, under ideal circumstances, enhances use of findings without jeopardizing quality or credibility; that stakeholders involved in the process sometimes develop an appreciation for and greater acceptance of evaluation, and gain skills associated with systematic inquiry; and that the process produces affective changes (such as feelings of self-worth) in participating stakeholders.

Weaknesses

Both proponents and critics have raised critical questions and concerns in regard to participatory evaluation. As Brisolara (1998) noted,

> One of the most frequent and serious charges leveled against PE [participatory evaluation] is that PE violates a long-held evaluation principle (or tradition) by forsaking an objective-as-possible stance for what some see as an inevitable slide into the stance of relativism. (p. 34)

Another common criticism of participatory forms of evaluation relates to technical quality. Some of the approach's detractors assert that technical quality is diluted by placing many decisions about evaluation methods and procedures into the hands of "nonexperts," thereby substantially reducing external credibility. On this point, Brisolara (1998) argued that "quality is maintained [by evaluators] through means similar to those adopted by their non-participatory colleagues—namely, evaluators remain responsible for ensuring the quality of methods and evaluation activities, and their role as evaluation expert is central to their function" (p. 36).

Participatory evaluation brings risks and challenges that are different from those of more traditional approaches, and its viability in practice (for example, when it comes to engaging large numbers of stakeholders) also has been brought into question (King, 1998, 2004, 2005). This concern about viability is especially salient when participatory evaluation is seen as threatening by some stakeholders (usually those with extant decision making authority and power) and when there is considerable stakeholder disengagement and apathy (Ryan et al., 1998). Other concerns have to do with particular aspects of usefulness—for example, how to define and identify program impacts and how to define the evaluator's role in sharing or giving away authority over decisions about designing, conducting, analyzing, and reporting on evaluations. Finally, Brisolara (1998) argued that participatory evaluation is "more of an implementation strategy . . . or community development 'dressed up' . . . than an evaluation approach" (p. 35).

Summary

In this chapter we have defined and discussed two eclectic evaluation approaches. An evaluator using an eclectic evaluation approach selectively borrows concepts and methods from the full range of legitimate evaluation models and approaches to meet the evaluation needs of particular stakeholder groups. Although the evaluation literature contains a wide range of eclectic evaluation approaches, this chapter concentrated on Patton's utilization-focused approach and Cousins's participatory approach. Among the unique characteristics of Patton's approach is its central focus on effectively addressing the intended uses of an evaluation for a selected group of intended users. Cousins's approach also is focused on serving stakeholders, but it is not restricted to serving a predefined group. Both approaches look to the stakeholders for direction in selecting evaluation questions and methods, and both require that evaluations meet the standards of the evaluation profession. The main difficulties with both approaches are in sustaining the involvement of stakeholders and ensuring that empowered stakeholders do not compromise the evaluation's rigor. Interested readers may also wish to pursue the referenced readings on Owen's evaluation forms approach (2004) and the W. K. Kellogg Foundation's approach (Council on Foundations, 1993; Millett, 1995, 1996).

REVIEW QUESTIONS

1. What features of Patton's utilization-focused evaluation approach led us to classify it as eclectic?
2. What features of the participatory evaluation approach led us to classify it as eclectic?
3. How did Patton conceive of the users to be served by a utilization-focused evaluation?
4. What are the advance organizers of the utilization-focused approach to evaluation?
5. What are the main elements of the evaluator's role in a participatory evaluation?
6. To what extent does Patton require that a utilization-focused evaluation serve socially valuable aims, such as combating social problems?
7. What is Patton's position in regard to the necessity of utilization-focused evaluations' fulfilling the requirements of the Joint Committee's program evaluation standards, and what do you think Patton would cite as the most important standard or criterion for judging a utilization-focused evaluation?
8. What does Patton mean by the active-reactive-adaptive method of utilization-focused evaluation?
9. What do we cite as the main limitations of participatory evaluation?

Group Exercises

This section is designed to support group discussion of key issues addressed in this chapter. Each exercise summarizes a particular case, then provides instructions for the group's analysis of and response to the case. After your group's members have read an exercise, engage in discussion to arrive at your group's response to the particular assignment.

Exercise 1

The managing director of a chain of sporting goods stores extending throughout New Zealand recently faced a stormy special meeting of shareholders as an outcome of a dismal forecast of annual profits for the ensuing three years. Among her responses to other demands from the meeting, she agreed that a completed survey and an assessment of the organization's functions would occur forthwith, and that a report would be furnished first to the board and then to shareholders at another special meeting within nine months. Those at the meeting also made it abundantly clear that the managing director's job would be on the line if the assessment findings indicated that she had been remiss in her leadership, decision making, and vision for the organization.

The managing director pondered her options. She could view the exercise simply from a financial perspective, enlisting the aid of the organization's firm of accountants, supported by the organization's auditors. But she knew that this would inevitably lead to drastic cost cutting, including a further reduction in staff (particularly at the middle-management level), a situation difficult to contemplate because a similar ploy three months earlier had led to falling staff morale at all stores. Or she could try to examine the real causes of the reduction in sales, however revelatory and painful this might be. Although economic factors clearly would emerge, she knew that such matters as poor and inadequate advertising, diminishing staff morale, suspect staff employment methods and training, and problems with business and associated program planning and execution were organizational weaknesses. These elements, she knew, would emerge from any examination of the organization.

She wisely decided to employ a reputable group of evaluators who could call on the assistance of financial experts as required.

Imagine that your group has been selected to conduct this evaluation. Your initial discussions with the managing director indicate that methodological pluralism will be essential to produce a sound and ethical report with its accompanying recommendations.

As a group, it is your task to discuss ways and means of convincing the managing director of the importance of pursuing the study based on an eclectic approach, given the wide range of concerns that exist. In the course of your deliberations, select either the utilization-focused or participatory approach and justify your choice. Then outline the presentation you would make to the managing director. Give special attention to the matter of engaging stakeholders.

Exercise 2

As this chapter has shown, Patton emphasized the practical implications of alternative (and multiple) evaluation approaches. The user must be given useful information and must be collaboratively involved in program assessment, with the evaluator acting as a guide and mentor.

Have one member of your group identify a program evaluation study known from personal experience or from published material. Then, as a group, construct a three-by-six matrix to compare the approach employed in the identified evaluation with Patton's utilization-focused evaluation approach. The column headings should be "Comparative Factors," the name of the identified evaluation, and "Utilization-Focused Evaluation." The row headings should be "Advance Organizers," "Purposes," "Sources of Questions," "Questions," "Strengths," and "Weaknesses." Fill in the cells of the matrix based on the one member's knowledge of the identified evaluation and the whole group's understanding of Patton's utilization-focused approach.

Exercise 3

Considering that participatory evaluation essentially gives the client group control of the selection of evaluation questions plus significant influence in choosing methods and interpreting findings, how can the evaluator ensure that the evaluation meets the following Joint Committee standards: U5 Relevant Information, U8 Concern for Consequences and Influence, P6 Conflicts of Interests, A2 Valid Information, and E2 Internal Metaevaluation?

Suggested Supplemental Readings

Alkin, M. C. (1985). *A guide for evaluation decision makers*. Thousand Oaks, CA: Sage.

Alkin, M. C. (1995, November). *Lessons learned about evaluation use*. Paper presented at the annual meeting of the American Evaluation Association, Vancouver, British Columbia, Canada.

Alkin, M. C. (2011). *Evaluation essentials: From A to Z*. New York, NY: Guilford Press.

Council on Foundations. (1993). *Evaluation for foundations: Concepts, cases, guidelines, and resources*. San Francisco, CA: Jossey-Bass.

Cousins, J. B. (1996). Consequences of researcher involvement in participatory evaluation. *Studies in Educational Evaluation*, *22*, 3–27.

Cousins, J. B. (2001). Do evaluator and program practitioner perspectives converge in collaborative evaluation? *Canadian Journal of Program Evaluation*, *16*, 113–133.

Cousins, J. B. (2003). Utilization effects of participatory evaluation. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 245–265). Norwell, MA: Kluwer.

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*, 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., & Whitmore, E. (1998). Framing participatory evaluation. In E. Whitmore (Ed.), *Understanding and practicing participatory evaluation* (pp. 5–23). New Directions for Evaluation, no. 80. San Francisco, CA: Jossey-Bass.

Cronbach, L. J., & Associates. (1980). *Toward reform of program evaluation*. San Francisco, CA: Jossey-Bass.

Cullen, A. E. (2009). *The politics and consequences of stakeholder participation in international development evaluations*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Cullen, A. E., & Coryn, C.L.S. (2011). Forms and functions of participatory evaluation: A review of the empirical and theoretical literature. *Journal of MultiDisciplinary Evaluation*, *7*(16), 32–47.

Cullen, A. E., Coryn, C.L.S., & Rugh, J. (2011). The politics and consequences of including stakeholders in international development evaluations. *American Journal of Evaluation*, *32*, 345–361.

Daigneault, P.-M., & Jacob, S. (2009). Toward accurate measurement of participation: Rethinking the conceptualization and operationalization of participatory evaluation. *American Journal of Evaluation*, *30*, 330–348.

Davis, H. R., & Salasin, S. E. (1975). The utilization of evaluation. In E. L. Struening & M. Guttentag (Eds.), *Handbook of evaluation research* (pp. 621–665). Thousand Oaks, CA: Sage.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Thousand Oaks, CA: Sage.

King, J. A. (1998). Making sense of participatory evaluation practice. In E. Whitmore (Ed.), *Understanding and practicing participatory evaluation* (pp. 57–67). New Directions for Evaluation, no. 80. San Francisco, CA: Jossey-Bass.

King, J. A. (2005). Participatory evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 291–294). Thousand Oaks, CA: Sage.

Millett, R. (1995). *W. K. Kellogg Foundation cluster evaluation model of evolving practices*. Battle Creek, MI: W. K. Kellogg Foundation.

Millett, R. (1996). Empowerment evaluation and the W. K. Kellogg Foundation. In D. M. Fetterman, A. J. Kaftarian, & A. Wandersman (Eds.), *Empowerment evaluation: Knowledge and tools for self-assessment and accountability* (pp. 65–76). Thousand Oaks, CA: Sage.

Owen, J. M. (2004). Evaluation forms: Toward an inclusive framework for evaluation practice. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 356–369). Thousand Oaks, CA: Sage.

Owen, J. M., & Rogers, P. J. (1999). *Program evaluation: Forms and approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (1980). *Qualitative evaluation methods*. Thousand Oaks, CA: Sage.

Patton, M. Q. (1982). *Practical evaluation*. Thousand Oaks, CA: Sage.

Patton, M. Q. (1994). Developmental evaluation. *Evaluation Practice*, *15*, 311–319.

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2003). Utilization-focused evaluation. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 223–244). Norwell, MA: Kluwer.

Patton, M. Q. (2004). The roots of utilization-focused evaluation. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 276–292). Thousand Oaks, CA: Sage.

Patton, M. Q. (2005). Utilization-focused evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 429–432). Thousand Oaks, CA: Sage.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Rodriguez-Campos, L. (2005). *Collaborative evaluations: A step-by-step model for the evaluator*. Tarmac, FL: Lumina Press.

Rossi, P. H., & Freeman H. E. (1993). *Evaluation: A systematic approach* (5th ed.). Thousand Oaks, CA: Sage.

Russon, C. (2005). Cluster evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 66–67). Thousand Oaks, CA: Sage.

Sanders, J. R. (1997). Cluster evaluation. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook* (pp. 396–404). Thousand Oaks, CA: Sage.

Scriven, M. (1980). *The logic of evaluation*. Inverness, CA: EdgePress.

Stufflebeam, D. L. (1983). The CIPP model for program evaluation. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 117–141). Norwell, MA: Kluwer.

Weiss, C. H. (1983). The stakeholder approach to evaluation: Origins and promise. In A. S. Bryk (Ed.), *Stakeholder-based evaluation* (pp. 3–14). New Directions for Program Evaluation, no. 17. San Francisco, CA: Jossey-Bass.

Whitmore, E. (Ed.). (1998). *Understanding and practicing participatory evaluation*. New Directions for Evaluation, no. 80. San Francisco, CA: Jossey-Bass.

CHAPTER 10

BEST APPROACHES FOR TWENTY-FIRST-CENTURY EVALUATIONS

This chapter provides a consumer report analysis and evaluation of four quasi-evaluation approaches; two improvement- and accountability-oriented approaches; two social agenda and advocacy approaches; and one eclectic approach. The chapter is intended to help evaluators and evaluation clients assess a selected set of broadly representative evaluation approaches and choose an approach that best meets both their particular needs and the professional standards of the evaluation field.

The approaches we chose for comparative analysis and evaluation are, in the quasi-evaluation category—objectives-based evaluation, the Success Case Method, the case study approach, and the experimental and quasi-experimental approach; in the improvement- and accountability-oriented category—the context, input, process, and product (CIPP) model and consumer-oriented evaluation; in the social agenda and advocacy category—constructivist evaluation and responsive or stakeholder-centered evaluation; and in the eclectic category—utilization-focused evaluation.

These approaches are applicable to program evaluations; representative of the different categories of legitimate evaluation approaches, as defined in this book; widely referenced in the professional literature (with the exception of the Success Case Method, which is a relative newcomer); and likely to be used extensively—advisedly or not—beyond 2014. In contrasting and evaluating the nine approaches, we have aimed to help evaluators and their clients critically appraise these approaches' particular strengths and weaknesses and potential utility before choosing among them.

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **Our rationale for selecting this chapter's nine particular evaluation approaches for systematic analysis and evaluation**
- **Ratings of the nine approaches against the utility, feasibility, propriety, accuracy, and evaluation accountability requirements of the Joint Committee on Standards for Educational Evaluation's *Program Evaluation Standards* (2011)**
- **The methodology and the metaevaluation checklist employed to produce our ratings of the nine approaches**
- **Our stated qualifications for rating the nine evaluation approaches, as well as caveats in regard to our conflicts of interest**
- **A comparison of 2007 and 2014 ratings of eight of the nine selected approaches (based, respectively, on the 1994 and 2011 editions of the Joint Committee's *Program Evaluation Standards*)**
- **Differences between the two sets of standards that may account for the modest differences observed between the 2007 and 2014 ratings**
- **The nine approaches' relative strengths, weaknesses, and utility**

Selection of Approaches for Analysis

We selected these particular approaches, rather than some of the other legitimate evaluation approaches referenced in preceding chapters, because (1) we sought balance in evaluating representative approaches in each category; (2) we needed to keep the approaches assessed to a manageable number; (3) these approaches, except for the Success Case Method, have been widely applied; and (4) we have had positive experiences in providing our students with detailed instruction on these particular approaches. We included the Success Case Method because of its unique focus on discovering program strengths, and because client groups, especially in industry, have welcomed the approach—for example, as a means of preserving and strengthening programs that may be unduly threatened for termination.

Our selection of the nine approaches reflects our experience-based preferences, but it should not be construed as an exclusion of the other legitimate approaches described in previous chapters. Indeed, thinking eclectically, it can be beneficial to incorporate aspects of the other approaches in applications of the selected nine.

Clearly some of the approaches not assessed in this chapter are worthy of consideration by evaluators and their clients. These include, especially, the cost study, value-added assessment, connoisseurship and criticism, accreditation and certification, deliberative democratic, and participatory approaches. We encourage readers who want an analysis and evaluation of approaches not included in our chosen set to consider selecting approaches of interest to them and then applying this chapter's methodology—as detailed in end-of-chapter notes—to analyze and evaluate the chosen approaches.

Methodology for Analyzing and Evaluating the Nine Approaches

Each of us independently rated the nine selected evaluation approaches on each of the thirty standards in the Joint Committee's *Program Evaluation Standards* (2011) by judging whether the approach endorses each of six key features of each standard. These ratings were produced using a detailed metaevaluation checklist, with six checkpoints for each of thirty standards. The first author determined these checkpoints by conducting a content analysis of the standards document.[1] Then together we resolved differences between our ratings and judged each approach's adequacy on each standard, depending on the number of checkpoints met: 6 = excellent, 5 = very good, 4 = good, 2–3 = fair, and 0–1 = poor. Scores for the approach on each of the five categories of standards (utility, feasibility, propriety, accuracy, and evaluation accountability) and overall were then determined according to systematic procedures.[2]

Our Qualifications as Raters

The first author's ratings were based on his knowledge of standards produced by the Joint Committee (1981, 1994, 2011); his many years of studying the various evaluation models and approaches; his personal acquaintance and collaborative evaluation work with authors and

leading proponents of all nine assessed approaches (Robert Brinkerhoff, Donald Campbell, Lee Cronbach, Egon Guba, Michael Patton, Michael Scriven, Robert Stake, and Ralph W. Tyler); and his experience in seeing and assessing how all of these approaches have worked in practice. He chaired the Joint Committee during its first thirteen years and led the development of the Joint Committee's first editions of both *The Program Evaluation Standards* (Joint Committee, 1981) and *The Personnel Evaluation Standards* (Joint Committee, 1988).

The second author's ratings were based on his teaching of the full range of approaches within the context of a doctoral-level evaluation theory course, as well as in doctoral-level experimental and quasi-experimental design, cost analysis, survey research, metaevaluation, and meta-analysis courses, as part of Western Michigan University's Interdisciplinary PhD in Evaluation program (which he directs); his national and international applications of a number of the approaches; and his collaborations with such authors as Tom Cook, Bradley Cousins, E. Jane Davidson, Daniel Stufflebeam, as well as Scriven, Patton, and Brinkerhoff.

Conflicts of Interest Pertaining to the Ratings

This chapter's ratings represent our informed but still personal judgments of the nine reviewed approaches. Both of us, as authors and raters, had conflicts of interest that readers should consider as they review the chapter's findings. Because the first author developed the CIPP approach, he had a conflict of interest in rating that model as one of the nine approaches. Also, readers may wish to consider, as a possible conflict of interest, that Scriven—author of the consumer-oriented approach—was the major professor for the second author's doctoral studies.

Our only controls designed to address conflicts of interest were that we produced independent sets of ratings prior to agreeing on final ratings, followed a rigorous rating protocol based on a metaevaluation checklist keyed to the 2011 Joint Committee program evaluation standards, and set forth in detail the rating procedures that followed.

One way for readers to assess the impact of our conflicts of interest on the produced ratings is independently to replicate this chapter's ratings of the subject nine approaches. Accordingly, readers could use the methodology underlying this chapter's ratings, including the detailed metaevaluation checklist that we ourselves used. We think this could be a highly instructive exercise, especially for evaluation students.

Standards for Judging Evaluation Approaches

We employed the 2011 Joint Committee program evaluation standards because they have been carefully thought through; were professionally developed by a highly credible and broadly representative Joint Committee; are widely accepted and applied, for example by the American Evaluation Association; and are nationally accredited by the American National Standards Institute—and especially because they are focused on ensuring that program evaluations are useful, feasible, ethical, accurate, and accountable.

The Rating Tool

Our ratings of the nine approaches are based not directly on the 2011 Joint Committee program evaluation standards, but on Stufflebeam's Program Evaluations Metaevaluation Checklist (2011b). The Joint Committee has neither reviewed nor sanctioned this checklist to confirm that it is a correct representation of the third edition of its 2011 *Program Evaluation Standards*.

Nevertheless, we stand behind this checklist as a carefully and systematically developed assessment tool. The first author determined the checklist's contents by carefully analyzing the contents of *The Program Evaluation Standards* (Joint Committee, 2011) and reflecting on his experience over the years in developing previous editions of that standards document plus helping with the development of the 2011 edition. His refinement and finalization of the checklist reflected his application of its initial versions in three metaevaluations conducted for the National Science Foundation and four for the national government of India, plus his use of the checklist in teaching a graduate-level practicum on metaevaluation. We are confident that the employed metaevaluation checklist will stand up to professional scrutiny, and, indeed, we invite independent assessments of the checklist.

Findings

The ratings of our selected nine approaches are shown in Figure 10.1. The approaches are listed in order of judged overall merit within the categories of quasi-evaluation, improvement- and accountability-oriented, social agenda and advocacy, and eclectic approaches.

The figure contains a synthesis of our judgments of each approach against the 180 checkpoints contained in the metaevaluation checklist we applied (six checkpoints for each of thirty standards). As already stated, that checklist was developed based on a content analysis of the Joint Committee's *Program Evaluation Standards* (2011). In the following discussion of the table we use the convention of referencing ratings for excellent, very good, good, fair, and poor by their abbreviations—E, VG, G, F, P.

The CIPP, utilization-focused, and constructivist evaluation approaches earned overall ratings of very good, whereas the other six approaches were judged good, overall. These results suggest that all nine approaches are at least minimally acceptable for evaluating programs but that all of them have room for improvement. The top-rated CIPP (85.68 percent–VG), utilization-focused (71.64 percent–VG), and constructivist (71.28 percent–VG) evaluation approaches provide strong options for evaluators who prefer either an improvement- and accountability-oriented approach, an eclectic approach, or a social agenda and advocacy approach. It is noted that the objectives-based (43.13 percent–G) and experimental and quasi-experimental (51.34 percent–G) options are least favorable for meeting the full range of the 2011 Joint Committee program evaluation standards, because they scored low in the good range. Also, it is worth considering that the responsive or stakeholder-centered (66.28 percent–G) and consumer-oriented (65.86 percent–G) approaches scored high in the range of good ratings.

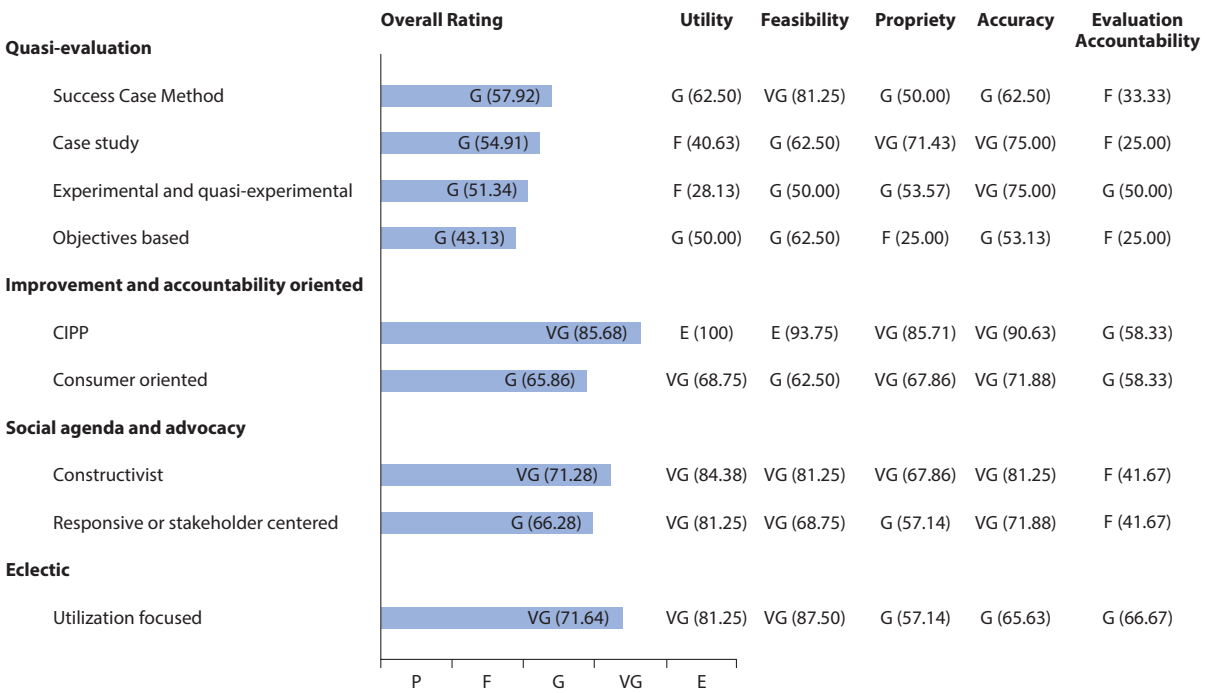

Figure 10.1 Strongest Program Evaluation Approaches Within Types in Order of Compliance with *The Program Evaluation Standards*[3]

Note: The scale ranges in the figure are P = poor, F = fair, G = good, VG = very good, and E = excellent.

Utility Ratings

If one is looking to maximize an evaluation's utility, the CIPP approach is the clear choice, with its utility rating of 100 percent–E. Other particularly good choices in the utility category are the constructivist (84.38 percent–VG) responsive (81.25 percent—VG) and utilization-focused (81.25 percent–VG) approaches. The experimental and quasi-experimental (28.13 percent–F) approach is the lowest rated in this category. Also, the objectives-based (50.00 percent–G) and case study (40.63 percent–F) approaches scored relatively low in the utility category and are not recommended when one wants to maximize an evaluation's utility and impact, including provision of systematic feedback during a program's development and operation. The consumer-oriented (68.75 percent–VG) approach's rating at the lower end of the very good range reflects this approach's emphasis on insulating summative evaluations from stakeholder involvement and influence and also its stronger emphasis on providing end-of-program, summative feedback than on providing feedback throughout a program's process. The utilization-focused (81.25 percent–VG) approach's high rating is lower than the CIPP model's rating against the utility standards, in part because of its selectiveness in serving only targeted stakeholders and potentially not serving some other right-to-know audiences.

Feasibility Ratings

In general, all nine approaches had ratings of at least good in the area of feasibility: CIPP (93.75 percent–E), utilization focused (87.50 percent–VG), success case methods and constructivist (81.25 percent–VG), responsive or stakeholder centered (68.75 percent–VG), consumer oriented (62.50 percent–G), case study (62.50 percent–G), objectives based (62.50 percent–G), and experimental and quasi-experimental (50.00 percent–G). The objectives-based approach's rating of good in the area of feasibility is consistent with its long tradition of widespread use. The experimental and quasi-experimental approach was judged to be the least feasible of the nine approaches, owing especially to its rigid application of separate treatments, holding treatments constant, and requiring randomization of subjects. Overall, the nine approaches were judged to be quite feasible for application.

Propriety Ratings

The top-rated approaches in the propriety category were the CIPP (85.71 percent–VG), case study (71.43 percent–VG), constructivist (67.86 percent–VG), and consumer-oriented (67.86 percent–VG) approaches. The CIPP approach's highest rating here is reflective of its explicit requirement for addressing and meeting the full range of 2011 Joint Committee program evaluation standards, including propriety requirements for advanced evaluation contracts, acknowledgment of and attention to the evaluator's conflicts of interest, control of report preparation and editing, and full and frank disclosure of findings to rightful audiences. In general, all four of these approaches have a strong ethical orientation, as defined in the propriety standards.

Somewhat lower but still quite acceptable ratings on propriety were obtained for the responsive or stakeholder-centered (57.14 percent–G), utilization-focused (57.14 percent–G),

experimental and quasi-experimental (53.57 percent–G), and Success Case Method (50.00 percent–G) approaches.

The rating for the objectives-based (25.00 percent–F) approach was in the lower end of the range of fair ratings, largely because this approach focuses almost exclusively on program goals; does not require validation of the goals through historical analysis and needs assessment; and is inadequate in addressing such propriety considerations as advance contracting, control of a staff's conflict of interest in evaluating its own program, and engagement of and provision of feedback to the full range of stakeholders throughout the evaluation process.

Based on the obtained ratings in the area of propriety, we would not recommend the use of the objectives-based approach. We believe the other eight approaches have basically satisfactory provisions for propriety. Moreover, we believe that applications of these approaches definitely could be strengthened by thoughtfully addressing the requirements of all of the Joint Committee's propriety standards (2011).

Accuracy Ratings

In the accuracy category, the top-rated approaches were the CIPP (90.63 percent–VG), constructivist (81.25 percent–VG), experimental and quasi-experimental (75.00 percent–VG), case study (75.00 percent–VG), consumer-oriented (71.88 percent–VG), and responsive or stakeholder-centered (71.88 percent–VG) approaches. The other three approaches received somewhat lower but still acceptable ratings in this category: utilization focused (65.63 percent–G), Success Case Method (62.50 percent–G), and objectives based (53.13 percent–G). In general, all nine approaches give evidence of at least satisfactory attention to matters of rigor. The comparatively low rating for the objectives-based approach is largely due to its narrow focus on main effects related to the developer's goals and little attention to context and side effects.

Evaluation Accountability Ratings

In the (new) evaluation accountability category, none of the nine approaches received particularly high ratings: utilization focused (66.67 percent–G), CIPP (58.33 percent–G), consumer oriented (58.33 percent–G), experimental and quasi-experimental (50.00 percent–G), constructivist (41.67 percent–F), responsive or stakeholder centered (41.67 percent–F), Success Case Method (33.33-F), objectives based (25.00 percent–F), and case study (25.00 percent–F). We think these findings point to a need both to strengthen the rated evaluation approaches in the area of evaluation accountability and to ask the Joint Committee to consider revising the External Metaevaluation evaluation accountability standard to improve its applicability.

It is noteworthy that the Joint Committee added the evaluation accountability group of standards to the 2011 edition of the standards because evaluators have, by and large, performed poorly in documenting, assessing, and securing independent metaevaluations of their evaluations. The relatively low ratings of the nine selected approaches in the evaluation accountability category reported here seem to support the need for standards, strengthened approaches, and

better evaluator performance pertaining to the three standards of evaluation accountability: Evaluation Documentation, Internal Metaevaluation, and External Metaevaluation.

Evaluators may justifiably disagree, however, with the External Metaevaluation standard's requirement that the evaluator engage and ensure the quality of an external metaevaluation. Arguably it is more appropriate for the evaluator to recommend that the client do this and then cooperate with an external metaevaluator, not by overseeing and controlling his or her work (which would entail a conflict of interest for the evaluator) but by supplying the external metaevaluator with needed information.

Comparison of 2007 and 2014 Ratings

In general, the 2014 ratings shown in Figure 10.1 are somewhat lower than the ratings recorded in a similar figure in the 2007 edition of this book (Stufflebeam & Shinkfield, 2007). The disparity is evident in Table 10.1, which contrasts overall 2007 and 2014 ratings given to the eight approaches that were assessed in the 2007 edition. (Note that the recently developed Success Case Method was not rated in the 2007 edition.)

The ratings in 2007 were determined by Stufflebeam and Shinkfield, whereas the 2014 ratings were arrived at by this book's authors—Stufflebeam and Chris Coryn. The Pearson product-moment correlation coefficient between the 2007 and 2014 overall ratings in Figure 10.1 was $r = 0.85$. This reflects considerable agreement, despite the changes in the 2011 edition of the Joint Committee's *Program Evaluation Standards* and the differing pairs of raters. A relatively high degree of consistency was observed between the rank order of approaches in the 2007 and 2014 ratings (Spearman's rank correlation coefficient was $\rho = 0.87$). Moreover, five of the eight approaches received the same general ratings (good or very good) in both studies. Overall ratings for the other three approaches dropped from very good in the 2007 study to good in the 2014 study.

The lower ratings of overall merit in this edition seem partially due to changes in the 2011 edition of the Joint Committee's *Program Evaluation Standards*. In particular, this edition added the evaluation accountability category. Figure 10.1 reveals that six of the nine evaluation

Table 10.1 Comparison of 2007 and 2014 Ratings of Eight Evaluation Approaches

Evaluation Approach	2007 Rating	2014 Rating
Case study	81.00%-VG	54.91%-G
Experimental and quasi-experimental	56.00%-G	51.34%-G
Objectives based	62.00%-G	43.13%-G
CIPP	92.00%-VG	85.68%-VG
Consumer oriented	84.00%-VG	65.86%-G
Constructivist	81.00%-VG	71.28%-VG
Responsive or stakeholder centered	84.00%-VG	66.28%-G
Utilization focused	86.00%-VG	71.64%-VG

approaches received lower ratings in this category than in the other four categories of standards. Another possibility for the lower ratings in our 2014 study is that the 2011 revised standards may be more demanding than their 1994 predecessors.[4]

Issues Related to the 2011 Program Evaluation Standards

We understand that proponents of particular evaluation approaches might take issue with our application of the 2011 Joint Committee standards as the criteria for judging the different approaches. We expect this is especially so in the cases of low ratings.

Clearly, some of the 2011 Joint Committee standards are at odds with what proponents of certain evaluation approaches consider to be best evaluation practice. For example, the requirement in *The Program Evaluation Standards* for detailed, preordinate planning of evaluation procedures clearly is contrary to requirements for flexibility in the social agenda and advocacy approaches. Authors of many of these approaches recommend an emergent, developing, evolving process for evaluation planning and implementation. Also, the 2011 standards place a pervasive, strong emphasis on a culturally sensitive and pluralistic approach to evaluation, including, it seems, empowerment of all stakeholders to exercise substantial control over evaluation planning, operations, and reporting.

Although the social agenda and advocacy evaluation approaches are congenial to stakeholder involvement and influence, proponents of other approaches, especially the experimental and quasi-experimental, CIPP, and consumer-oriented evaluation approaches, see delegation to stakeholders of influence and control over evaluation matters as a threat to an evaluation's independence, rigor, and credibility.

To some extent, certain negative judgments of evaluation approaches based on the 2011 Joint Committee standards may be viewed as unwarranted or at least questionable, with the resulting interpretation that some aspects of the new standards themselves may be unacceptable to credible experts in the evaluation field. In any case, readers should exercise circumspection in viewing ratings based on the 2011 Joint Committee program evaluation standards that are decidedly lower than the corresponding ratings appearing in Stufflebeam and Shinkfield's 2007 edition of this book.

Overall Observations

In rounding out this chapter, offering commentary on salient features of each approach is in order. Our comments are grouped in terms of this book's four categories of legitimate evaluation approaches.

Quasi-Evaluation Approaches

The Success Case Method and case study approaches both scored overall in the middle of the range of good ratings. Of the two, the Success Case Method was shown to be more useful—62.50 percent versus 40.63 percent—than the straight case study approach. We think

this is so because the former is directly oriented to discovering particular program strengths that may be hard to detect and whose detection might help program staff preserve and strengthen a program that otherwise could be headed for termination.

In contrast, the case study evaluator seeks more to give a rich account of a program than to issue judgments related to saving or strengthening a program. In regard to propriety, the case study approach has the clear edge over the Success Case Method—71.43 percent versus 50.00 percent—partially because the latter approach starts out with a built-in bias toward finding and reporting successes and is more geared to serving a program director and staff rather than the full range of program stakeholders. Both approaches are weak in meeting the standards of evaluation accountability

The experimental and quasi-experimental and objectives-based approaches overall rated at the lower end of the range of good ratings. We believe both approaches have limited applicability in the broad range of program evaluation assignments.

The comparatively low overall rating given to the experimental and quasi-experimental approach resulted especially from its rating of fair for utility. For many evaluation assignments, this approach would not provide program staff members with continuing feedback for program improvement; and for program evaluation applications in the field, the approach often proves to be impractical, vulnerable to political problems, and not cost effective. Its rating of 75.00 percent in the accuracy category, which although in the very good range is lower than one might have expected, is due more to its narrow focus on a few dependent variables and lack of information on context and process than to the quality of the obtained, focal outcomes.

The overall rating of good for the objectives-based approach is somewhat misleading, because it scored at the very bottom of the range of good ratings. This poor showing reflects the approach's narrow focus on objectives, provision of only terminal information, and lack of attention to unanticipated outcomes. For most program evaluation assignments, evaluators are advised to seek a better approach than either the experimental and quasi-experimental or objectives-based evaluation approach.

The Success Case Method and case study approaches are our methods of choice in the quasi-evaluation category of evaluation approaches. In comparison to other approaches, however, they are nonetheless quite weak choices.

Improvement- and Accountability-Oriented Approaches

The improvement- and accountability-oriented approaches rated slightly better overall than the quasi-evaluation, social agenda and advocacy, and eclectic approaches.

The CIPP approach's generally high ratings in Figure 10.1 reflect its comprehensiveness in assessing all stages of program development and all aspects of a program; serving the full range of stakeholders; employing multiple quantitative and qualitative methods; providing for formative and summative uses of findings; being oriented to both program improvement and accountability; addressing all thirty of the Joint Committee's 2011 program evaluation standards; requiring at least internal metaevaluation; and, especially, being grounded in advance agreements keyed to stakeholder needs and professional standards for evaluations. The approach's relatively

low rating in regard to evaluation accountability reflects the approach's disagreement with the External Metaevaluation standard, which thrusts the evaluator into a conflict of interest situation by requiring the evaluator rather than the client to select and engage an external metaevaluator and to oversee his or her work.

The consumer-oriented approach's particular strength is in providing clients and evaluators with independent, unimpeachable assessments of programs, services, and developed products. Although the approach is not strongly suited to internal evaluations for improvement, it complements such approaches with an outsider, expert view that becomes important when products and services are put up for dissemination. This approach depends on a highly skilled evaluator who strongly guards independence and separation from program personnel. Paradoxically, the approach depends on program personnel for much of the information needed for the evaluation, which tends to be the approach's Achilles' heel. As a cautionary note, a high degree of evaluator independence from program personnel can discourage the extensive amount of stakeholder support that the consumer-oriented evaluator often needs. This psychological distance also can discourage program personnel from using external evaluation findings, but it can be reassuring to external audiences—especially those who pay for the program or use its products and services. This approach's relatively high rating on utility is due not to a strong impact on the actions of program personnel during program implementation, but to the high degree of credibility consumers external to a program place on independent evaluations.

Social Agenda and Advocacy Approaches

The two social agenda and advocacy approaches generally scored well, definitely ahead of the quasi-evaluation approaches. Constructivist evaluation is a well founded, mainly qualitative approach to evaluation that systematically engages interested parties to help conduct both the divergent and convergent stages of evaluation. The constructivist evaluator strongly advocates for the least powerful and most economically disadvantaged among program stakeholders. The approach tends to be utopian and thus unrealistic, which is acknowledged by its creators. Also, its provision for ongoing negotiation with a wide range of stakeholders—through hermeneutic and consensus-building processes—engenders difficulty in reaching closure under a framework of multiple values and multiple realities. Nevertheless, this approach earned quite acceptable ratings in utility, feasibility, propriety, and accuracy. Its rating on evaluation accountability was low but, as with the other rated approaches, it reflects the evaluation field's historical poor performance in documenting and evaluating evaluations.

The responsive or stakeholder-centered approach received quite acceptable ratings overall and for utility, feasibility, propriety, and accuracy. Like the constructivist evaluation approach, it received a low rating in the area of evaluation accountability. In contrast to the consumer-oriented approach, with its emphasis on independence, the responsive approach engenders close collaboration between the evaluator, program personnel, and other stakeholders, resulting in easier access to needed information and stakeholders' better acceptance, support, and use of an evaluation. This approach has the advantage of systematically informing and assisting ongoing development and operations. It is also strong in searching for unintended consequences.

Its comparatively low rating in the propriety category reflects its lack of provision for advance formal contracting for evaluation; lack of focus on meeting published, professional standards for sound evaluations; and quite weak approach to identifying and addressing conflicts of interest. Aside from these points, the approach is credible in the area of propriety, especially with its strong orientation to evenhandedly engaging and serving the full range of stakeholders.

Eclectic Approaches

Finally, utilization-focused evaluation is a ubiquitous, umbrella approach to evaluation. It received an overall rating of very good, ratings of very good in utility and feasibility, and ratings of good in the evaluation accountability, accuracy, and propriety categories. Its main objective is to get evaluation findings used and accordingly rates high on utility. As noted earlier, the approach's rating on utility, which is lower than one might expect, is due to its mission of serving exclusively a predefined set of stakeholders—leaving the possibility that some right-to-know audiences might not be served, plus the difficulty of sustaining the involvement of targeted stakeholders. This approach also rates high on feasibility, because stakeholders are invited both to help ensure that the study will fit well in their program's environment and to choose such elements as evaluation questions and reporting schedules and modes with which they are comfortable. The approach emphasizes efficiencies, including using existing information and incorporating insider knowledge into the evaluation process. Especially, the approach directs the evaluator to foster use and impacts of findings by involving and addressing the evaluation needs of a narrowly defined group of intended users. Utilization-focused evaluation stands in contrast to the notion that has been popular in evaluation circles that an evaluator must strive to identify and address a wide range of evaluation questions that are of interest to all possible program stakeholders. In general, the utilization-focused approach places an evaluation's focus where the intended users of evaluation findings want it—which enhances ease of use and evaluation impact. The relatively low rating on propriety, although still in the good range, is largely due to the approach's possible limiting of service to only a subset of right-to-know stakeholders. The good rating on accuracy is not higher in large part due to the approach's penalty received for not necessarily resulting in a printed report. Accordingly, a utilization-focused evaluator might not produce a written report that documents and justifies an evaluation's technical merit, his or her adherence to professional standards, and the chain of reasoning that led to the evaluation's conclusions. Nevertheless, the approach is viewed as strong in respect to exercising rigor in collecting and analyzing both qualitative and quantitative information and especially in respect to getting evaluation findings used.

The Bottom Line

Based on the ratings presented in this chapter, evaluators and their clients can choose from an array of strong, creditable evaluation approaches. When assessed against professional standards for program evaluations, the best approaches are the CIPP, consumer-oriented, constructivist, utilization-focused, and responsive or stakeholder-centered approaches. All of

these approaches are recommended for consideration in program evaluations. We believe that better alternatives to objectives-based and experimental and quasi-experimental approaches typically can be found.

Nevertheless, it is essential to stress that even the strongest of the nine approaches have considerable room for improvement. Clearly, all of the approaches should be strengthened in the vital area of evaluation accountability. We think Figure 10.1 can be profitably used to identify and target areas in each approach for needed improvements.

Also, we encourage all who are so inclined to either adapt or apply as is the metaevaluation checklist we used to make their own assessments of evaluation approaches that interest them. Such approaches could include ones described in the literature or ones being applied in the field.

Summary

This chapter was prepared in the format of a consumer report to provide evaluators and evaluation clients with a systematic analysis and associated judgments of nine selected evaluation approaches drawn from four categories of such approaches. The approaches assessed were, for the quasi-evaluation group, the Success Case Method, case study, experimental and quasi-experimental, and objectives-based approaches; for the improvement- and accountability-oriented group, the CIPP and consumer-oriented approaches; for the social agenda and advocacy group, the constructivist and responsive or stakeholder-centered approaches; and for the eclectic group, the utilization-focused approach.

Using Stufflebeam's Program Evaluations Metaevaluation Checklist, we rated each approach against six checkpoints for each of thirty standards for sound evaluations. The employed metaevaluation checklist had been developed based on a content analysis of the 2011 Joint Committee program evaluation standards. Those standards spell out requirements for evaluations in five areas: utility, feasibility, propriety, accuracy, and evaluation accountability. Each approach was rated overall and for each of the five categories of standards on a five point scale: excellent, very good, good, fair, and poor.

In advance of presenting the ratings of the nine selected approaches, we reported on our rationale for selecting the nine approaches for analysis, commented on our qualifications to rate the approaches, acknowledged our conflicts of interest in producing the ratings, noted what we had done to mitigate the conflicts, and documented the methodology and referenced the checklist employed to produce the ratings.

Following are some of the chapter's key findings:

- Based on overall ratings against the thirty Joint Committee standards, the best approaches were, in order of rated merit (highest first), the CIPP (very good), utilization-focused (very good), constructivist (very good), responsive or stakeholder-centered (good), and consumer-oriented (good) approaches.
- Based on overall ratings against the thirty Joint Committee standards, the weakest approaches were, in order of rated merit (highest first), the Success Case Method (good),

case study (good), experimental and quasi-experimental (good), and objectives-based (good) approaches.

- Ratings of the best approaches against standards of utility, feasibility, propriety, and accuracy ranged from good to excellent, whereas ratings of the four weakest approaches ranged from fair to very good.
- In general, ratings of the nine approaches against the evaluation accountability standards were lower than their ratings on utility, feasibility, propriety, and accuracy. The ratings on evaluation accountability were (from lowest to highest) case study (fair), objectives-based (fair), Success Case Method (fair), constructivist (fair), responsive or stakeholder centered (fair), experimental and quasi-experimental (good), consumer oriented (good), CIPP (good), and utilization focused (good).
- The lower ratings in the evaluation accountability category may be due to this category's only recent (2011) inclusion in the Joint Committee's *Program Evaluation Standards* and to a requirement that the evaluator (rather than the client) be responsible for obtaining an external metaevaluation. Regarding the latter point, many evaluators could reject this requirement because of the threat to independence when an evaluator controls evaluation of her or his own evaluation.

Overall, we concluded that evaluators and their clients can choose from an array of creditable evaluation approaches, especially in the improvement- and accountability-oriented, social agenda and advocacy, and eclectic categories. The weakest approaches were judged to be in the quasi-evaluation category.

REVIEW QUESTIONS

1. This question is, in part, preparation for the other review questions in this section. Having perused Chapters 6 through 9, write a brief paragraph identifying the main intentions of each of these evaluation approaches:
   a. Success Case Method
   b. CIPP
   c. Consumer oriented
   d. Responsive and stakeholder centered
   e. Constructivist
   f. Utilization focused
2. Why have we rated the objectives-based approach comparatively low on compliance with the Joint Committee's 2011 *Program Evaluation Standards* with respect to the utility and propriety standards?

3. Are you able to support the very good rating of the experimental and quasi-experimental approach in the accuracy category, considering the fair rating in the utility category? Why or why not?
4. Justify or refute the assertion that there is, among the evaluation approaches discussed in this chapter, "an increasingly balanced quest for rigor, relevance, and justice."
5. What are the inherent weaknesses of social agenda and advocacy evaluations, and also of improvement- and accountability-oriented studies?

Group Exercises

Exercise 1

This exercise will have greater benefit for group members if they prepare their responses in advance of the meeting. If the exercise is undertaken thoroughly, it should serve two main purposes: first, further study of the six approaches listed in the first review question, and second, a review of Chapter 3 on standards for program evaluation as well as a practical application of the standards.

This exercise focuses on the Success Case Method, CIPP, consumer-oriented, responsive or stakeholder-centered, constructivist, and utilization-focused approaches to program evaluation. Each group member is allocated one of these six. If there are more than six group members, a particular approach will be allocated more than once.

Refer to Figure 10.1 and the associated end-of-chapter notes, and study the procedures underlying the ratings used in the figure (at first glance, they may appear complicated, but in reality the method is quite straightforward). Using the same methodology, each group member should give a rating for his or her allocated approach in each of the areas of utility, feasibility, propriety, accuracy, and evaluation accountability, as well as an overall rating. Group members will need to download and copy the Program Evaluations Metaevaluation Checklist from www.wmich.edu/evalctr/checklists or from the book's Web site at www.josseybass.com/go/evalmodels. Clearly, subjectivity and degree of experience in program development and evaluation will play a part in group members' ratings. However, knowledge of the allocated evaluation approach and the exact nature (by definition and example) of each of the thirty standards will provide very useful parameters for decision making.

As a group, discuss

- The proximity of each member's ratings to ours
- Possible reasons for any wide divergences
- The benefits of knowing and using the 2011 Joint Committee program evaluation standards and the associated Program Evaluations Metaevaluation Checklist

Exercise 2

Discuss the ramifications of the following statement: "If evaluators ignore the likely conflicts in purposes for an evaluation, the program evaluation is probably doomed to fail." As a group, discuss whether you agree or disagree with this statement, and explain the reasons for your agreement or disagreement.

Notes

1. The particular tool used to rate each evaluation approach was Stufflebeam's Program Evaluations Metaevaluation Checklist (2011b), available at www.wmich.edu/evalctr/checklists, and also at the Jossey-Bass Web site that supports the use of this book, www.josseybass.com/go/evalmodels.
2. Ratings for each category of the 2011 Joint Committee standards (utility, feasibility, propriety, accuracy, and evaluation accountability) were obtained by summing the following products: 4 × number of excellent ratings, 3 × number of very good ratings, 2 × number of good ratings, and 1 × number of fair ratings. The sum was then divided by this product: 4 × number of standards in the category. Judgments of each approach's strength in satisfying each category of standards were then determined according to percentages of the possible quality points for the category of standards as follows: 92–100 percent = excellent, 67–91.99 percent = very good, 42–66.99 percent = good, 17–41.99 percent = fair, and 0–16.99 percent = poor. The final percentage scores were obtained by multiplying the initial decimal point score obtained for each category and overall by 100. The five equalized percentage scores were then summed and divided by 5. The result was then judged by comparing it to the total maximum score, 100 percent. Each approach's overall merit and merit for each category of standards were judged as follows: 92–100 percent = excellent, 67–91.99 percent = very good, 42–66.99 percent = good, 17–41.99 percent = fair, and 0–16.99 percent = poor. After each of us rated each of the nine selected evaluation approaches following the method just described, we systematically reviewed and discussed our different sets of ratings and reached consensus judgments wherever we found discrepancies in ratings of each approach—overall and for utility, feasibility, propriety, accuracy, and evaluation accountability. In this process, we also sought to reach a "plus" or "minus" determination for those checkpoints to which one or the other of us had assigned a question mark. We ultimately reached the consensus judgments presented in Figure 10.1.
3. The ratings are based on the 2011 Joint Committee program evaluation standards. We arrived at the ratings through use of Stufflebeam's checklist (2011b) keyed to the those standards, which is available both from this book's Web site at www.josseybass.com/go/evalmodels and the Evaluation Center's Web site at www.wmich.edu/evalctr/checklists. The checklist essentially reflects a systematic content analysis of the 2011 standards and consists of 180 discrete checkpoints (48 for the eight utility standards, 24 for the four feasibility standards, 42 for the seven propriety standards, 48 for the eight accuracy standards, and 18 for the three evaluation accountability standards).

   Ratings were completed for each of the domains of the program evaluation standards (utility, feasibility, propriety, accuracy, and evaluation accountability) and overall (Joint Committee, 2011). We first independently rated each of the nine approaches against each of the thirty standards by deciding whether or not the approach met each of the six checkpoints. Assignments for each checkpoint were assigned a plus sign if the approach fulfilled the requirement, a minus sign if not, or a question mark if it was unclear whether or not the approach embraced or addressed the particular requirement. In using the ratings to arrive at judgments of each approach—for each category and

overall—only "pluses" were scored. Subsequently we jointly reviewed our ratings, discussed and resolved discrepancies, attempted to reach determinations for checkpoints that had been assigned a question mark, and finally produced Figure 10.1.

4. Another useful contrast to the current and former (Stufflebeam & Shinkfield, 2007) ratings is Stufflebeam's widely cited *Evaluation Models* monograph (2001b), in which a similar method was employed to evaluate the relative merits of various evaluation approaches against the second edition of *The Program Evaluation Standards* (Joint Committee, 1994).

Suggested Supplemental Readings

American Evaluation Association. (2004). *Guiding principles for evaluators*. Washington, DC: Author. Retrieved from http://www.archive.eval.org/Publications/GuidingPrinciples.asp

Joint Committee on Standards for Educational Evaluation. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

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

Stufflebeam, D. L. (2011). *Program Evaluations Metaevaluation Checklist*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/checklists

Stufflebeam, D. L., & Shinkfield, A. J. (2007). *Evaluation theory, models, and applications*. San Francisco, CA: Jossey-Bass.

PART THREE

EXPLICATION OF SELECTED EVALUATION APPROACHES

The chapters in Part Three are designed to help readers develop a firm grasp of six approaches to evaluation and their applicability. In Chapters 11 through 16 we provide in-depth information about each of those six approaches—experimental and quasi-experimental design evaluation; case study evaluation; the context, input, process, and product (CIPP) evaluation model; consumer-oriented evaluation; responsive or stakeholder-centered evaluation; and utilization-focused evaluation.

CHAPTER 11

EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGN EVALUATIONS

The experimental and quasi-experimental design approach to program evaluation is intended to produce unbiased conclusions about a program's effectiveness. The typical experimental design–based program evaluation involves random assignment of individuals, groups, or other units to one or more conditions (such as treatment, alternative treatment, zero-treatment control, treatment as usual, placebo control, wait-list control, or attention control); applying a special treatment to one group and none (or an alternative treatment) to another group; holding treatment conditions constant throughout an evaluation; and ultimately assessing and contrasting groups' posttreatment performance on one or more outcome variables of interest. It should be noted that the term *randomized experiments,* as discussed in this chapter, is used synonymously to describe variants that are sometimes used in some disciplinary traditions (for example, randomized clinical trials, randomized controlled trials, or true experiments).

Chapter Overview

In this chapter we discuss randomization as the key element in experimentally based evaluations. We define randomized experimental and quasi-experimental designs and discuss them as limited but useful evaluation approaches that should be a part of every program evaluator's repertoire (and at least be understood at a conceptual level). We define experimental design's important but limited role in causal investigations, especially by acknowledging that prospective, randomized experimentation

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **The central tenets and logic of experimental and quasi-experimental design evaluations**
- **Uses and misapplications of experimental designs in program evaluations**
- **Conditions required for conducting experimental studies**
- **Edward Suchman's seminal contributions to the experimental method**
- **Exemplars of large-scale experimental studies**
- **Guidelines for designing and executing experiments**
- **Key alternatives to randomized controlled experimental designs**
- **The strengths and weaknesses of experimental and quasi-experimental evaluations**

is centrally important to the cause-and-effect paradigm but not to the paradigm covering studies that retrospectively search out causes of observed effects. We outline the philosophy, principles, and practices of an early, dominant figure by discussing what Suchman in the 1960s termed "the scientific approach to evaluation." We then fast-forward to the early twenty-first century and discuss contemporary conceptualizations and principles of randomized field trials and experimental and quasi-experimental designs for evaluation, although many, many such advances—most of which have been premised on and derived from the medical model of cause and effect—occurred during the period between the 1960s and the 2000s. In addition to Suchman's work (1967), particularly influential writings on the experimental and quasi-experimental design approach were Campbell and Stanley's *Experimental and Quasi-Experimental Designs for Research* (1966), Cook and Campbell's *Quasi-Experimentation: Design and Analysis for Field Settings* (1979), and Shadish, Cook, and Campbell's *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (2002). The last of these is widely considered the definitive text on experimental and quasi-experimental designs for manipulating and examining the effects of social interventions.

In addition, we summarize different real-world adaptations of experimental designs as reviewed by Nave, Miech, and Mosteller (2000) and others. Then, drawing from the conceptualizations of Suchman (1967), Boruch (2003), and Shadish et al. (2002), in particular (though many others have made substantial contributions), and from some of our own (based on our experiences implementing these types of studies as well as our analysis of such approaches in the historical and contemporary literature), we present guidelines for planning and executing experimental and quasi-experimental design evaluations.

Basic Requirements of Sound Experiments

Random assignment of experimental subjects to treatment and control (or alternative treatment) conditions is the sine qua non of the true experiment. It is done to ensure that (1) groups are probabilistically equivalent (based on fair and equal chances of assignment to comparison groups and assuming sufficiently large samples), (2) posttreatment differences in outcomes are due only to comparison groups' different treatment experiences, and (3) an evaluator can estimate the statistical probability that observed differences are not due to chance variation or other alternative explanations (Shadish et al., 2002). Assuming there are sufficient numbers of people, groups, or other units in each condition, an experimental design strengthens the likelihood that groups are equivalent on all measured and unmeasured variables that might influence their response to the program being investigated. Any differences in outcomes that emerge between treatment and control groups following treatment implementation can, if the required assumptions and conditions have been satisfied, be attributed to differences in treatments.

In general, applications of experimental design require substantial resources, a high level of methodological expertise, strong political commitment, random assignment of subjects to conditions, sustained cooperation of experimental subjects, and relatively long periods of control over treatment implementation and data access and collection.

Prospective Versus Retrospective Studies of Cause

Coryn and Hobson (2011) noted the following in regard to inferences made from experimental and many types of quasi-experimental designs:

> For the majority of evaluations such inferences are about the effects of a given cause rather than questions about the cause of a given effect. Put simply, a cause is that which precedes or produces an effect, and an effect is the difference between what occurred in the presence of a (presumed) cause and what occurred in its absence (i.e., counterfactual reasoning; Rubin, 1974, 2005). Additionally, this view of causation is premised on manipulable causes that can be deliberately varied and that can generate reasonable approximations of the physically impossible counterfactual. Based on this logic, three conditions are necessary for causal inference (Shadish et al., 2002): (a) temporal precedence—that cause precedes effect, (b) covariation—that cause and effect vary together, and (c) absence of alternative causes—that no other plausible explanations can account for an observed treatment-outcome covariation. (p. 32)

Notably, experiments require careful prospective planning and design (that is, prior to the delivery of a treatment or program), whereas most other types of designs, including many types of quasi-experiments, are often retrospective in nature (in other words, designed and conducted after a treatment or program has already occurred). However, by comparison with randomized, prospective, variable-manipulating experiments, studies to determine causes of observed effects are also critically important in the realm of studying causes. This is evident in such post hoc investigations as an in-depth case study to determine the root causes of a child's reading disability; a coroner's autopsy to determine the one or more causes of a death; a laboratory's studies to help a physician diagnose a patient's difficulty in swallowing food; an engineering firm's forensic investigation to determine why a bridge collapsed; historical and correlational analyses that established the link between lung cancers and smoking tobacco; a causal, detailed historical analysis to determine why Napoleon lost the battle of Waterloo; law enforcement/judicial, DNA-assisted proceedings to determine responsibility for a murder; and an epidemiological investigation to locate the underlying cause or causes of illness associated with eating cantaloupes from a certain vendor. As is evident in such familiar post hoc investigations, a randomized experiment is but one of many varying approaches to investigating and reaching conclusions about causes. The wide array of evaluators and other investigators who study and reach conclusions about causes collectively require both the cause-and-effect and the effect-and-cause paradigms to guide their investigations.

Uses of Experimental Design

Studies of the applications of randomized experimental design reveal a fairly large family of related configurations for studying program effects. Such studies provide clear examples of sound experiments producing valuable assessments of causes and effects. However, the studies also indicate that many field-based experimental studies are poorly implemented.

Boruch (2003) documented uses of experimental design for evaluating programs in a wide range of service areas, including employment; criminal justice; health care; cultural enrichment programs for children; preschool, elementary, and secondary education; distance education; and AIDS reduction. Nave et al. (2000) reviewed seven randomized field trials of educational programs. These are noteworthy for the range of designs employed to apply the principles of experimental design.

Spybrook (2008), in her investigation of studies funded by the U.S. Department of Education's Institute of Education Sciences, found that a majority of those studies were underpowered. More recently, Christie and Fleischer (2010) reviewed the research designs used and reported in studies published in North American evaluation journals, finding that randomized experiments are relatively rare in certain areas of evaluation practice despite the attention that such designs frequently receive in the scholarly literature.

Randomized Controlled Experiments in Context

Opportunities to meet the requirements of randomized experiments are quite limited. This is especially true in such service fields as education, transportation, and social services and in innovative development projects in which a premium is placed on creativity, trial-and-error exploration, and continual feedback for improvement. The randomized experimental design approach seems to work best when an investigator can muster strong control of experimental variables. Prominent examples are testing seeds, fertilizers, pesticides, and cultivation practices in agriculture; new drugs in the pharmaceutical industry; innovative procedures in medicine; and different behavioral stimuli in experimental psychology.

The Limited Applicability of Randomized Controlled Experiments

Arguments used against the employment of experiments usually are of three kinds. First, it is often contended that a group or groups deprived of a new and supposedly better treatment are at a decided disadvantage (especially when funds have been targeted to provide equitable assistance to all members of a certain needs-based group, such as a school district's disadvantaged elementary school students). The second argument is that it is often impossible to keep groups separate and free from a wide range of contaminating factors. In the real world of schools, for example, it is often difficult to arrange and sustain treatment and control groups and convince administrators and parents and other stakeholders that the benefits of such procedures are worth the difficulties of keeping groups separate and sustaining their different treatment experiences. Clearly, programs that seek continuous improvement and innovative breakthroughs, and that need continuous feedback on what is or is not working well, can be stifled when placed under strict controls required by a true experimental design. That being said, however, Shadish et al. (2002) have advised that interventions with unsubstantiated beneficial effects and/or negative side effects, should not be widely disseminated until evidence of their effectiveness has been adequately established. Lastly, the third argument against use of experimental design is that the goal of many naturalistic evaluations of innovative, field-based programs is to monitor and

provide feedback to help improve the program, while also compiling in-depth qualitative and quantitative evidence of its evolution, quality, costs, and outcomes. Such summative evidence is used both for accountability purposes and to understand the program's process, costs, and consequences, which generally are beyond the scope of most experiments. In such applications, so the argument goes, the program's staff members benefit from continuous feedback for improvement, and the program's constituents ultimately receive an in-depth summative report based on the detailed record of the program's implementation, costs, and outcomes, with the possibility of receiving conclusions about program success or lack thereof that are beyond reasonable doubt.

Boruch (2003) acknowledged that experimental design is not applicable to many program evaluation situations. Certainly it is not a panacea. We disagree sharply with the representatives of the U.S. Department of Education who have often publicly cited randomized controlled trials as the gold standard of program evaluation and mandated the approach's—nearly—exclusive use for federally funded evaluations of educational programs through the Education Sciences Reform Act of 2002 (U.S. Department of Education, 2003). In 1951 R. A. Fisher, a pioneer of experimental design, warned that his experimental design approach should not be considered the only, or most appropriate, method of inquiry for all situations, which is what the Institute of Educational Sciences of the U.S. Department of Education (2003) did—a position it later retracted.

Since the mid-1960s, U.S. government agencies concerned with education cyclically have imposed and subsequently retreated from mandates to use randomized experiments to evaluate federal education projects. Repeatedly, the retractions have come when the unfeasibility of using rigid designs to evaluate dynamic, improvement-oriented projects has become clear. This pattern of what we judge to have been misguided federal leadership constitutes, in our view, significant waste of federal resources for evaluating needed educational improvement projects.

An Example of Misapplication of the Experimental Approach

As one personal example of this rather harsh judgment, in the early 1970s this book's first author served on a metaevaluation panel for monitoring and assessing a federally supported, $5 million experimental design evaluation of a large urban school district's federally supported Emergency School Assistance Act (ESAA) project. The project was targeted to help the district meet the needs of a high influx of students newly arrived due to soldiers' families' being stationed at a nearby military installation in the context of the nation's military buildup to fight the Vietnam War. This buildup suddenly had brought thousands of new students to several school districts. ESAA had been established to help the districts meet the needs of these students. At the federal government's mandate, this evaluation (and others like it) had to be conducted strictly in accordance with the requirements of randomized controlled experimental design. Senator Patrick Moynihan, in consultation with Professor Donald Campbell of Northwestern University and in support of Campbell's advocacy of an experimenting society, had authored a federal mandate requiring that ESAA be evaluated in accordance with the requirements of true experiments.

The district's contracted evaluation group identified the treatment as ESAA funds, identified matched pairs of schools in the district, and randomly assigned ESAA funds to

one member of each pair. The evaluators then set out to measure the effects of the ESAA treatment (added funds) by examining a number of outcome measures, especially standardized achievement test results. At the first opportunity to search for differential effects, it became clear that there were no significant differences between the experimental and control groups' assessed outcomes. Moreover, diagnostic investigation revealed that the district's superintendent had, to avoid dissension in the district, allocated to the control schools other district funds in amounts that approximated the ESAA allocations to the experimental schools. Thus, essentially there were no significant differences in treatments (money allocated to each control and treatment school), and, not surprisingly, none in assessed educational achievement outcomes for the two sets of schools.

Subsequently, the evaluators abandoned the mandated experimental arrangements and proceeded with an attempt to conduct an in-depth case study. Unfortunately, their team lacked experience and expertise in the area of case study evaluation. Ironically, another evaluation group—with strong credentials in qualitative methodology—originally had submitted a proposal projecting that the mandated randomized controlled experimental design approach would not work and that a competently conducted case study approach could provide valuable information on how the schools used the ESAA funds and with what discernible results. Unfortunately, their proposal had been rejected without serious consideration of its rationale, practicality, and potential to produce useful information. Following the failed attempt to assess the treatment (consisting only of allocated federal money) by means of a randomized experiment, the contracted evaluation team regrouped and struggled to learn and apply sound case study methods.

As seen in this example, although experimental design is valuable in a limited sphere, there are many instances in the context of evaluation in which experimental design is not needed, is inappropriate, is not feasible, is potentially wasteful of resources, or may even be counterproductive. These include especially the different stages of formative evaluation—for example, needs assessment, evaluation of program proposals, process evaluation, exception reporting, cost analysis, and monitoring for quality assurance. In such situations, randomization and holding treatments constant probably could constitute undue interference, be irrelevant, and/or be counterproductive. In these cases, options preferable to experimental design could include rigorous application of any of a number of observational and analytical approaches, including especially the case study approach; the Success Case Method; self-reports followed by site visits by visiting teams; connoisseurship; the context, input, process, and product (CIPP) evaluation model; consumer-oriented evaluation; responsive evaluation; goal-free evaluation; and utilization-focused evaluation (Stufflebeam, 2001b).

Alternative Approaches for Addressing Cause-and-Effect Questions

In the sphere of cause-and-effect evaluations, experimental design is only one of a host of applicable methods. Scriven (2005a, 2009a; also see Cook, Scriven, Coryn, & Evergreen, 2010) made this clear when he argued persuasively that causal inferences can be obtained and strongly defended from a range of approaches other than randomized, comparative

experiments. He posited that the correct gold standard of cause-and-effect program evaluations is not conclusions from mandated randomized experiments but rather conclusions beyond reasonable doubt, whatever the employed method. Such conclusions are obtainable from rigorous application of a wide range of approaches, including, but extending far beyond, experimental design. (We illustrated these previously when we distinguished between the cause-and-effect and effect-and-cause paradigms.) Some of the wide range of approaches available for investigating cause are evident in the following examples: studies of germs and remedies applied to germ-free animals in gnotobiotic laboratories; epidemiological studies to determine causal agents in disease epidemics; diagnostic studies in medicine; diagnostic investigations of failures in buildings, automobiles, airplanes, and traffic control systems; DNA laboratory tests to determine paternity or criminal involvement; interrupted time-series studies; in-depth case studies; and regression discontinuity designs.

Contexts in Which Experiments Are Particularly Applicable

We debated whether to devote a complete chapter to experimental and quasi-experimental design, thereby possibly giving this approach undue significance, particularly by comparison with the other widely applicable approaches explored and discussed in Part Three. The main reason for our dilemma was that like many others (but not all) in the evaluation field, we hold some reservations about the utility and feasibility of implementing randomized controlled experiments in the dynamic worlds of education, social services, and other evaluative domains. The experimental approach's requirements for random assignment, control of treatments, and withholding of ongoing feedback for program improvement can inhibit study and improvement of a program's merit and worth (Stufflebeam, 2001b). Conversely, it is possible to support the application of randomized experiments and quasi-experiments to evaluation when they are used appropriately to address cause-and-effect questions. Thus, despite the strong and often justified criticisms of experimental design, we believe that the approach should not be disregarded. Moreover, many important principles from the experimental paradigm should be considered requisite knowledge for evaluators.

In certain well-defined circumstances, randomized controlled experimentation has a place in program evaluation and may elicit valuable information for decision making, especially in the realm of large-scale, highly funded policy evaluation.

For example, Finkelstein et al. (2011) recently conducted a randomized experimental study of a sample of uninsured, low-income adults in Oregon. The study's subjects were selected by lottery from the larger group of low-income, potential Medicaid applicants and given the opportunity to apply for Medicaid. (Because demand for Medicaid insurance exceeded available funds, control subjects were not considered to be unduly deprived of insurance.) In the year after random assignment, the treatment group was about 25 percent more likely to have insurance than the control group that was not selected. Advantages for the treatment group included substantively and statistically significantly higher health care use (including primary and preventive care as well as hospitalizations), lower out-of-pocket medical expenditures and medical debt (including fewer bills sent to collection), and better self-reported physical and mental health than the control group (Finkelstein et al., 2011).

As this example shows, randomized controlled experimental design is a powerful approach to addressing cause-and-effect questions under real-world conditions, when circumstances warrant. Such circumstances may include a larger group of beneficiaries than can be served with available resources—making random selection and assignment of subjects ethically and politically acceptable; a treatment that can be differentially assigned to a randomized group of recipients, such as enrollment in an experimental entitlement program; reasonable control over who receives the treatment, such as with selection by lottery; low need to control beneficiaries' use of the treatment over time, such as when treatment beneficiaries choose to use treatment services based on their needs; and outcome data on treatment and control subjects that are normally collected, stored, and available for inspection, such as in government records and databases, plus experimental and control subjects who are willing to contribute self-reports. Though rare, the fact that such circumstances may exist makes randomized controlled experimentation a powerful, viable option to consider under appropriate conditions. Evaluators are advised to conduct rigorous evaluability studies (see M. F. Smith, 1989) before proceeding with experimental studies.

Suchman and the Scientific Approach to Evaluation

Suchman (1967) set the principles of experimental design within a broad view of policy evaluation and stressed the importance of taking account of the relevant social context. His seminal writings on the topic greatly influenced many other prominent evaluation theorists and methodologists, including Donald Campbell, Huey Chen, Thomas Cook, Lee Cronbach, Mark Lipsey, Peter Rossi, William Shadish, and Carol Weiss, among many others (also see Alkin & Christie, 2004). It is of note that Suchman made a clear demarcation between evaluation, which he equated with judgment, and evaluation research, which he denoted as judgment grounded in scientific research. For a time in the 1970s, *evaluation research* was a term much in vogue (for example, Weiss, 1972). It referred to an evaluation (that is, a judgment) based on empirical research and subject to criteria of sound research, especially reliability, validity, generalizability, and objectivity (also see Coryn, 2007b). The term *evaluation research* has largely since disappeared from the evaluation literature and predominately has been replaced by *evaluation* or *program evaluation*. Nevertheless, Suchman's early writings (1967) about evaluation research are clearly relevant to, and resonate with, the evaluation field as we know it today.

Suchman (1967) held that research scientists must base their conclusions primarily on scientific evidence. It follows that he believed that evaluation must be approached with the logic of the scientific method. His work and writings during the 1960s, however, emphasized the need to assess a program in relation to its practical setting. For this reason, he suggested specific criteria for assessing program success. His studies in the field of social sciences, particularly public health, made him keenly aware that evaluation research is attended by practical constraints. Moreover, he stated that evaluation researchers, in their attempts to expose desirable and undesirable consequences, must consider relevant values, especially those in conflict.

In a key contribution to the field, *Evaluative Research: Principles and Practice in Public Service and Social Action Programs* (Suchman, 1967), which was greatly influenced by Campbell

and Stanley (1963, 1966; also see Mark, Donaldson, & Campbell, 2011), Suchman stressed that evaluators should use whatever research techniques are available and appropriate to the circumstances and needs of a particular evaluation study. Although he believed the ideal study would adhere to the classical experimental model, he also stressed that, in reality, evaluation research projects usually involve some variation or adaptation of this model. To a large extent, according to Suchman's reasoning (1967), formulation of the objectives and design of an evaluation research study largely depends on who conducts the study and the anticipated use of outcomes.

Suchman (1967) was not alone in his belief that although evaluators are basically researchers, they must strike a balance between rigorous methodology and adapting to the situation in which they must function. Earlier writers who had advocated a similar approach to evaluation methodology included Klineberg (1955), James (1958), Herzog (1959), and Fleck (1961). Suchman differed from these writers by distinguishing clearly between evaluation and evaluation research. The former he referred to generally as "the process of making judgments of worth," whereas the latter he considered to be "those procedures for collecting and analyzing data which increase the possibility for proving rather than asserting the worth of some social activity" (Suchman, p. 62). One could deduce that by distinguishing evaluation from evaluation research, Suchman was placing the evaluator in a technical role and reserving the valuational interpretation role for the client, whom he often referred to as the "administrator."

When Suchman (1967) discussed the process of evaluation, he proposed a scientific approach grounded in logical positivism. He saw evaluation as a continuous social process, inherently involving a combination of basic assumptions underlying the activity being evaluated and the personal values of the study participants as well as of the evaluator. Evaluation, he maintained, must necessarily become a scientific process to take into account this intrinsic subjectivity, because it cannot be eliminated. With the development of models and approaches that espouse constructivism and postmodernism and whose proponents have not sought to eliminate subjectivity, many evaluators today would not embrace Suchman's focus on the precepts, standards, and methods of logical positivism and hypothetico-deductive research. It would be a decade after Suchman's untimely death in 1971 that intensive work on broader views of standards for evaluation would be undertaken, and some years thereafter before they were being accepted and applied. The development of standards was a watershed in the history of evaluation. These include the standards and guidelines reviewed in Chapter 3 and others keyed to experimental studies, including ethical requirements for experiments prescribed by the U.S. Federal Judicial Center and federal requirements placed on institutional review boards in the United States.

Viewing evaluation as a scientific process, Suchman (1967) maintained that the same procedures that we use to discover knowledge could be used to evaluate the degree of success in the application of this knowledge. He held strongly to the concept that by adopting the scientific method, an evaluator will produce findings that are more objective and of ascertainable reliability and validity. He viewed evaluation research as applied research and (with Ralph W. Tyler) saw its purpose as determining the extent to which a specified program is achieving the desired results. He said the results should be geared to helping administrators make sound

decisions about the program's future. Bearing in mind the dominant role of administrative criteria in determining a study's value, Suchman warned that the evaluator must be constantly aware of the potential utility of findings. This emphasis on the necessity of useful findings could give rise, he believed, to a very real problem for the evaluator. Because of strong vested interest, a program administrator might endeavor to control the evaluation, at least to the extent of defining the objectives of the program to be evaluated. To a far greater extent than the basic researcher, the evaluator could lose control of the area being investigated. Thus, in Suchman's view, it is not the concepts of research per se that make evaluation studies difficult, but rather the practical problems of adhering to these principles in the face of administrative considerations.

When he explored whether evaluation research was ready to play a more significant role in policymaking, Suchman (1967) concluded that it was not. Systematic analysis of the theoretical, methodological, and administrative principles underlying the evaluator's objectives and procedures was needed before positive and meaningful steps forward could be taken confidently. It would not be too far fetched to assume that over four decades ago, Suchman was assuming that guidelines were necessary for both evaluator and client. We believe that he would have welcomed the sets of guiding principles and standards that have since been established.

Suchman's Purposes and Principles of Evaluation

Suchman (1967) supported the purposes of evaluation enumerated by Bigman (1961):

- To discover whether and how well objectives are being fulfilled
- To determine the reasons for specific successes and failures
- To uncover the principles underlying a successful program
- To direct the course of experiments with techniques for increasing effectiveness
- To lay the basis for further research on the reasons for the relative success of alternative techniques
- To redefine the means to be used for obtaining objectives and even to redefine subgoals in light of research findings

These purposes, in Suchman's opinion (1967), suggested an intrinsic relationship between program planning and development, on the one hand, and evaluation, on the other. In effect, the procedures used to achieve these evaluation purposes must provide the basic information for designing and, if necessary, redesigning programs. Just as traditional research should lead toward increased understanding of basic processes, so should evaluation research "aim at an increased understanding of applied or administrative processes" (Suchman, p. 64). This emphasis on the importance of administration and its processes is quite relevant to any attempt to evaluate the success or otherwise of a program five years after conclusion of the original evaluation.

A principle that Suchman (1967) strongly endorsed is that different situations warrant different evaluation approaches, including different technical methods and criteria for measuring

the success in obtaining desired objectives. Looking at the assumption that an evaluation study may take several forms and also that the primary function of most studies is to help stakeholders design, develop, and operate programs, we again see Suchman drawing the distinction between evaluation and evaluation research. He considered evaluation as a goal, and evaluation research as a particular means of obtaining that goal.

We strongly stress that although Suchman (1967) believed that the use of scientific methodology needed a particular emphasis, he did not rule out the use of nonscientific methods. He acknowledged that in program design and implementation, many evaluation questions can be answered without research. Nevertheless, he maintained that if the basic requirements of evaluation research could be met—that is, underlying assumptions of objectives examined, measurable criteria developed specifically related to objectives, and a controlled situation instituted—then conclusions based on convincing research, and not just subjective judgment, would be the outcome.

Values and the Evaluation Process

A precondition of any evaluation study, Suchman (1967) maintained, is the presence of some activity whose objectives are assumed to have value. He defined value as "any aspect of a situation, event, or object that is invested with a preferential interest as being good, bad, desirable, undesirable, or the like" (p. 67). From this we may construe that values—as modes of organizing human activity based on principles that determine both the goals and implementation of programs as well as the means of obtaining these goals—are basic precepts of any evaluation study. Suchman argued that the evaluation process stems from and returns to the formation of values, as shown in Figure 11.1.

Despite a gap of decades, Suchman's philosophy (1967) that evaluation starts with a particular value (either explicit or implicit) and then proceeds to goal setting—that is, a selection among alternative goals—is in accord with some, but not all, current philosophies of evaluation. Goal-setting forces are necessarily in competition with each other for resources and effort. We cannot argue against criteria being selected to measure the attainment of goals or against the idea that the nature of the goals would determine some of the measures used. Even so, a number of defensible evaluation approaches are not narrowly focused on goals when it comes to selecting evaluative questions, criteria, and methods. Moreover, Suchman adopted a Tylerian approach (R. W. Tyler, 1942) when he required that the evaluation be used to ascertain the degree to which the operating program has achieved the predetermined objectives. Finally, based on this orientation, he sought a judgment as to whether goal-directed activities were worthwhile.

Figure 11.1 indicates that the act of judgment returns the activity to value formation. Suchman's concept (1967) of the cyclical movement of the evaluation process emphasizes the close interrelationship between evaluation and the value-laden nature of program planning and operation. As a result, there is the ever-present possibility of a conflict of values between the program administrator and the evaluator. In general terms, it can be said that values play a large role in determining the objectives of social science programs and that the evaluation process that exposes desirable and undesirable consequences of such programs must take into account societal values, especially conflicting ones.

Figure 11.1 Suchman's Evaluation Process

Source: Suchman, E. A. (1967). *Evaluative research: Principles and practice in public service and social action programs*. New York, NY: Russell Sage Foundation, 34.

Assumptions for Evaluation Research Studies

Suchman's main assumption (1967) for evaluation studies was that every program has some value for some purpose. It follows that the most identifying feature of evaluation research "is the presence of some goal or objective whose measure of attainment constitutes the main focus of the research problem" (Suchman, p. 71).

When a clear statement of the program objective to be attained has been explicated, the evaluation may be viewed as a study of change. The program to be evaluated constitutes the causal or independent variable, and the desired change is similar to the effect or dependent variable. Characterizing an evaluation study this way, Suchman (1967) postulated that the project may be formulated in terms of a series of hypotheses that state that activities A, B, and C will produce results X, Y, and Z (that is, descriptive causation—*that* A causes B; also see Coryn, Noakes, Westine, & Schröter, 2011; Shadish et al., 2002).

Objectives and assumptions of an evaluation study are closely tied when the following difficult questions need to be answered before a study commences: What kinds of changes are desired? What means will be used to effect these changes? What signs will enable the changes to be recognized? Before these questions can be addressed adequately, the evaluator must be able to diagnose the presence or absence of a social problem and its underlying value system and to define goals indicative of progress in ameliorating that condition.

Suchman (1967) outlined six questions that must be answered when formulating the objectives of a program for evaluation purposes and, indeed, the design of the study itself:

- What is the nature of the content of the objective (for example, change in knowledge, attitudes, or behavior)?
- Who is the target of the program (for example, large-scale or discrete groups)?
- When is the desired change to take place (for example, are there short-term or long-term goals or cyclical, repetitive programs)?
- Are the objectives unitary or multiple (for example, are the programs similar for all users or different for different groups)?
- What is the desired magnitude of the effect (for example, widespread or concentrated results)?
- How is the objective to be attained (for example, voluntary cooperation or mandatory sanctions)?

Many of the answers to these questions will require an examination of the assumptions underlying the stated objectives. Suchman (1967) saw it as the duty of an evaluator to challenge these assumptions if necessary, stressing that only then can the "scientific" label truly be applied to the evaluation process.

Suchman (1967) classified assumptions into two types: value assumptions and validity assumptions. Value assumptions pertain to the system of beliefs that determines what is good within a society or part of that society (Shadish, Cook, & Leviton, 1991). For example, a new educational program may be viewed favorably or unfavorably by various groups within a school district. The question the evaluator must answer before investigating the program is, What is success, and from whose point of view?

Validity assumptions are specifically related to program objectives. Such assumptions, for example, underlie the belief of educators that early elementary programs must be consonant with the home influences of each child. Suchman (1967) stressed that answers to all validity questions can never be discovered before a program is initiated. Administrators should call on their experience and skill to develop practical programs whose assumptions are clearly laid down. The task of the evaluator is then to prove or disprove the significance and defensibility of these assumptions.

It is obvious that Suchman (1967), with his emphasis on experimental designs, was well aware of the world around the evaluation researcher and the necessity of dealing with a program's constituents, with all their strengths, weaknesses, idiosyncrasies, and values.

Suchman's Categories of Evaluation

Suchman (1967) spoke of three categories of evaluation studies. Ultimate evaluation refers to the determination of the overall success of a program in relation to its statement of objectives. Preevaluative evaluation deals with intermediate problems (for example, development of reliable and valid explications of the problem, definition of goals, and perfection of techniques) that must be solved before ultimate evaluation may be attempted. Short-term evaluation involves seeking specific information about concrete procedures in terms of immediate utility.

Suchman (1967) maintained that evaluation research could be conducted in terms of different categories of effect, in addition to varying levels of objectives. These categories represent various criteria of success (or failure) by which a program is judged. He proposed five categories:

- *Effort*. Evaluations in this category have as their criteria of success measures of the quantity and quality of program activity that takes place. This is an assessment of input regardless of output. It indicates at least that something is being done to solve a problem.
- *Performance*. Criteria in this area measure results of effort rather than the effort itself.
- *Adequacy of performance*. This category refers to the degree to which effective performance is adequate by comparison with the total amount of need (according to defined objectives).
- *Efficiency*. Evaluation in this category addresses the question, Is the capacity of an individual, organization, facility, operation, or activity to produce results in proportion to the effort expended?
- *Process*. The purpose of this category is to investigate basic reasons underlying the findings. Suchman outlined four dimensions of an analysis of process: the attributes of the program; the population exposed to the program; the context within which the program occurred; and the different kinds of effects produced by the program (for example, multiple or unitary effects and duration of effects).

In summary, in discussing types and categories of evaluation, Suchman (1967) outlined a basic process to be followed in conducting an evaluation study. This process entails stating objectives in terms of ultimate, intermediate, or immediate goals; examining the underlying assumptions; and instituting criteria of effort, performance, adequacy of performance, efficiency, and process.

Methodological Aspects of Experimental Designs

Suchman (1967) characterized evaluation research as a special kind of applied research whose goal, unlike that of nonevaluation or basic research, is not to discover knowledge but to test the application of knowledge. Emphasizing its utility, Suchman posited that evaluation research should provide information for program planning, implementation, and development. Evaluation research, with its various experimental designs, also assumes the particular characteristic of applied research that allows predictions as to the outcome of investigation. Suchman said that recommendations made in evaluation reports are examples of predictions. Unlike basic, laboratory research, an evaluation study based on experimental design contains an array of variables over which the evaluator has little if any control. He stressed that in evaluation research, the observable and measurable variables are the phenomena of interest; the implemented program has as its goal the changing values of these measures. Whereas the underlying concept is of prime importance in basic, theory-testing research, such is not the case in applied research, of which evaluation research is a form.

Suchman (1967), inadvertently or otherwise, struck on the main limitation of the experimental approach in evaluation when he contended that program evaluation may have very little generalizability because an evaluation is applicable solely to the program being evaluated within its context (see also Campbell, 1986). In other words, uncertainty is always present in that a program that is effective in one situation very well may not be effective in another. Moreover, when problem-solving objectives of evaluation research are considered to give strength to administrative decision making for specific needs, further difficulties for external validity always present themselves.

The inherent difference between evaluation research and basic research studies, according to Suchman (1967) and earlier writers, is reflected in the form taken by the statement of the problem. Whereas pure research (that is, nonevaluation research) questions whether A is related to B and tests this relationship experimentally (under controlled conditions), applied research (evaluation research) questions whether A works effectively to change B and attempts to answer this question empirically by manipulating A and measuring the effect on B. In basic, nonevaluation research, crucial importance is given to an analysis of a process whereby A relates to B; in evaluation research, it is far less important to understand *how* A produced B (that is, a causal explanation; see also Coryn, Noakes, et al. 2011; Shadish et al., 2002). These days, evaluation emphasis must be given to the process that occurs between program initiation and findings. All five evaluation approaches that follow in Part Three (in Chapters 12 to 16) adhere to this concept, as do many modern conceptions of experimental design.

The differences between evaluation and nonevaluation research are not absolute, but may be considered to exist on a continuum. Suchman (1967) thought that an evaluator therefore should follow as closely as is practical the rules of scientific inquiry, but the evaluator must define and justify where and when these rules have to be adapted to reality.

Principles of Evaluation Research Design

Suchman (1967) presented a list of principles to be observed in laying out the design of an evaluation study; the main ones are as follows:

- A good design is one that is the most suitable for the purpose of the study; whether or not it is scientific is not an issue.
- There is no such thing as a single correct design; hypotheses can be studied by different methods using different designs.
- All research design represents a compromise dictated by many practical considerations.
- An evaluation research design is not a highly specific plan to be followed without deviation, but rather a series of guidelines to ensure that main directions are followed.

Chapter 6 of Suchman's 1967 book gives details of possible variations in evaluation research design, which interested readers may wish to investigate. Although Suchman believed that the ideal study would adhere to the classical experimental model, he also stressed that in reality, evaluation research projects usually use some variation or adaptation of this model. To a large

extent, the formulation of the objectives and design of an evaluation research study will depend on who conducts the study and the anticipated use of the findings. He emphasized that while designing a study, an evaluator must be aware that validity considerations are crucial.

According to Suchman (1967), the measurement of the effects of a program requires specification relating to four major categories of variables:

1. Component parts or processes of the program
2. The intended population and actual groups reached
3. Situational conditions within which the program occurred
4. Differential effects of the program

Suchman (1967) recognized that determining both the reliability and the validity of the criteria of effectiveness of these variables causes particular problems. For the most part, the evaluator measures not the phenomena being studied directly, but rather indexes of these phenomena (latent rather than direct manifestations). Two obvious problems present themselves. First, how does the evaluator decide on indicators for the criteria of achievement of program objectives? Second, how does the evaluator select from all possible indicators those to be used for a particular purpose?

In presenting ways to solve these problems, Suchman (1967) discussed aspects of the methodological concepts of reliability and validity at considerable length. In particular, he emphasized that evaluators should be aware of and endeavor to control the main sources of unsystematic variation in evaluation research:

- *Subject reliability*. Attitudes and behavior are affected by moods, fatigue, and so on.
- *Observer (evaluator) reliability*. Personal factors influence interpretation of a subject's responses.
- *Situational reliability*. Conditions of measurement produce changes in results that do not reflect true changes in the evaluand being studied.
- *Instrument reliability*. All of the preceding factors (combined) or specific aspects of the instrument itself (for example, poorly worded questions) affect reliability.
- *Processing reliability*. Coding errors, occurring randomly, lower reliability.

Validity presents a much broader problem than reliability because it refers not only to specific measures but also to the significance of the whole evaluation process. The validity of an evaluation study refers to the validity of its specific measures; it also refers to the theory underlying the hypotheses relating the evaluation's activities to its objectives. Suchman (1967) identified the following sources of bias leading to validity concerns in evaluation studies:

- *Propositional validity*: The use of incorrect or inappropriate assumptions or theories
- *Instrument validity*: The use of irrelevant operational indexes
- *Sampling validity*: Lack of population representativeness in the sample

- *Observer (evaluator) validity*: Introduction of a consistent bias based on personal bias or preconceived notions
- *Subject validity*: Habits and predispositions of subjects that introduce invalid biases
- *Administration validity*: Conditions of the study (for example, methods of data collection) that constitute a source of invalidity
- *Analysis validity*: Deliberate or unintended bias in analysis that causes invalidity

The differential effects of a program encompass what Suchman (1967) termed "unanticipated or unintended effects." Social phenomena are so complex and interrelated that to change one of its aspects becomes impossible without producing a series of connected changes, which Suchman termed "secondary effects." These secondary effects of a program can be particularly troublesome when the program is intended to be widely disseminated. Federally funded programs in education fall into this category just as much today as they did in Suchman's time. The evaluator and program administrator must therefore be wary of easy acceptance of secondary positive effects as justification for a program even if its intended objectives are not achieved; decisions concerning generalizability should take into account both intended and unintended effects. However, it is also important to understand that research is a learning process and that analysis of secondary effects (desirable or undesirable) is an integral part of this process.

Contemporary Concepts Associated with the Experimental and Quasi-Experimental Design Approach to Evaluation

Despite Suchman's profound influence (1967), most contemporary philosophical thinking about and methodology associated with the experimental and quasi-experimental design approach to program evaluation stem from Shadish et al.'s seminal 2002 book *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Modern experimental reasoning is premised on activity theory, whose key feature is deliberate variation of something so as to discover what happens to something else later—to arrive at presumed causes. Moreover, this approach is predominately premised on manipulable causes (for example, treatments, interventions, programs). Nonmanipulable events (such as a hurricane) and traits (such as gender) cannot be deliberately varied to determine their effect on something else (but they can still be studied as causal antecedents). A central task for all cause-probing studies, therefore, is to create reasonable approximations of the physically impossible counterfactual. As mentioned earlier in this chapter, "such inferences are about the effects of a given cause rather than questions about the cause of a given effect" (Coryn & Hobson, 2011, p. 32). Central to this reasoning is that (1) cause precedes effect, (2) cause and effect vary together, and (3) no other plausible explanations can account for an observed causal relationship (where a cause is that which produces an effect or result, an effect is the difference between what did happen and what would have happened or that which occurs as a result of something else, and a counterfactual is knowledge of what would have happened in the absence of a presumed cause).

Intuitively, the causal effect of an experimental treatment, *E,* over a control condition, *C,* for a particular unit and an interval of time from t_1 to t_2 is the difference between what would have

Table 11.1 Basic Counterfactual Logic

Subject	Y_E	Y_C	$Y_E - Y_C$
Subject 1	5	?	?
Subject 2	6	?	?
Subject 3	?	2	?
Subject 4	?	3	?
Subject 5	5	?	?
Subject 6	?	2	?
Subject 7	5	?	?
Subject 8	?	3	?
M	5.25	2.50	2.75

happened at time t_2 if the unit had been exposed to E initiated at t_1 and what would have happened at t_2 if the unit had been exposed to C initiated at t_1 (Rubin, 1974). Given that subjects cannot simultaneously be members of both E and C groups, only the average causal effect can be estimated, as illustrated in Table 11.1, where the question marks indicate responses that cannot be observed.

One of the major concerns among those who advocate the experimental and quasi-experimental design approach to evaluation are threats to validity. Threats to validity are reasons why an inference may be incorrect. Many threats to validity cannot be ruled out with design controls, either because the logic of design control does not apply (for example, because of inadequate construct explication) or because practical constraints prevent such controls from being applied (for example, there are intact groups). The major focus of design is, therefore, anticipating and reducing the number and plausibility of threats to validity. In contrast to Suchman (1967), Shadish et al. (2002) enumerated four major types of validity:

1. *Internal validity*: The validity of inferences about whether the relationship between two variables is causal
2. *Construct validity*: The degree to which inferences are warranted from the observed persons, settings, treatments, and cause-and-effect operations sampled within a study to the constructs that these samples represent
3. *External validity*: The validity of inferences about whether a causal relationship holds over variations in persons, settings, treatment variables, and measurement variables
4. *Statistical conclusion validity*: The validity of inferences about the covariation between two variables

Experimental and quasi-experimental design evaluators are primarily concerned with internal validity, though the other types of validity also are considered important. As noted earlier, threats to internal validity are reasons why inferences that the relationship between two variables is causal may be incorrect; these threats include, but are not limited to, the following (Shadish et al., 2002):

1. *Ambiguous temporal precedence*. Lack of clarity about which variable occurred first may yield confusion about which variable is the cause and which is the effect.
2. *Selection*. Systematic differences in respondent characteristics across conditions could cause the observed effect.
3. *History*. Events occurring concurrently with treatment could cause the observed effect.
4. *Maturation*. Naturally occurring changes over time could be confused with a treatment effect.
5. *Regression*. When units are selected for their extreme scores, they will often have less extreme scores on other variables, an occurrence that can be confused with a treatment effect.
6. *Attrition*. Loss of respondents to treatment or to measurement can produce artifactual effects if that loss is systematically correlated with conditions.
7. *Testing*. Exposure to a test can affect scores on subsequent exposures to that test, an occurrence that can be confused with a treatment effect.
8. *Instrumentation*. The nature of a measure may change over time or over conditions in a way that could be confused with a treatment effect.
9. *Additive and interactive threats*. The impact of a threat can be additive in relation to that of another threat or may depend on the level of another threat.

By definition, experimental designs rule out selection threats due to random assignment. In short, these types of designs are intended to strengthen causal inferences through the application of design elements. These elements include assignment (for example, random assignment, cutoff-based assignment); measurement (pretests, posttests, retrospective pretests, nonequivalent dependent variables); comparison groups (single nonequivalent comparison groups, multiple nonequivalent comparison groups, cohorts); and treatments (switching replications, removed treatments). In characterizing experimental and quasi-experimental designs it can be useful to use the notation shown in Table 11.2 to most efficiently identify potential (contextual) validity threats as well as to clearly explicate the design used for an evaluation.

The basic, two-group, posttest-only experimental design would be configured using the notation in Table 11.2 as follows (other common and less common types are explained in detail in Shadish et al. [2002]):

R	X	O_1
R		O_1

Notably, this design does not include a pretest observation. Pretest observations should be avoided if sensitization (that is, exposure to a measurement procedure that influences subsequent measurements) is likely; if a pretest cannot feasibly be gathered; or is a constant (as in studies of mortality in which all patients are alive at pretest), for example. Pretests are a necessity when attrition is likely—and it nearly always is—and as a means of increasing

Table 11.2 Common Notation for Experimental and Quasi-Experimental Designs

Notation	Definition
R	Random assignment
NR	Nonrandom assignment
O	Observation
X	Treatment
~~*X*~~	Removed treatment
X_+	Treatment expected to produce an effect in one direction
X_-	Conceptually opposite treatment expected to reverse an effect
C	Cut score
—	Nonrandomly formed groups
. . .	Cohort

statistical power by using them as covariates. Extending this design to include a pretest observation, which might be useful for exploring attrition, using the pretest as a covariate, and/or determining whether randomization was successful (that is, ensuring that groups were equivalent prior to assignment to the treatment or control condition), the design would be:

R	O_1	X	O_2
R	O_1		O_2

These designs are intended not only to reduce certain validity threats but also to address different types of questions. For example, the multiple-treatment design that follows could be used to address the relative effectiveness of one treatment (*A*) versus another (*B*) in comparison to a control, to compare specific components or parts of a treatment, or for dose-response studies (differing doses of the same treatment):

R	O_1	X_A	O_2
R	O_1	X_B	O_2
R	O_1		O_2

Quasi-experimental designs share all structural features with experimental designs except that units are not randomly assigned to conditions. One of the most common quasi-experimental designs used for evaluation, the untreated control group design with dependent pretest and posttest samples, would be notated as follows:

NR	O_1	X	O_2
NR	O_1		O_2

This design is similar to the randomized pretest-posttest design shown earlier except that units are not randomly assigned to treatment or control conditions. With this design, and nearly

all quasi-experimental designs, a selection bias is always present, but the pretest observation allows for determining the magnitude and direction of bias. An even more common design used for program evaluation is the one-group pretest-posttest design:

O_1	X	O_2

Here, the pretest provides only weak information concerning what might have happened to participants had the program not occurred.

Although the basic designs just described do not do justice to the complex options available for experimental and quasi-experimental designs for evaluation (for example, factorial designs), later in this chapter we will discuss the state of the art in quasi-experimental alternatives to experimental designs, including regression discontinuity designs, interrupted time-series designs, and other quasi-experimental designs that often approximate experimental studies.

Exemplars of Large-Scale Experimental and Quasi-Experimental Design Evaluations

Next, it is of interest to move beyond this theoretical conception of experimental design and see how the concept has worked out in practice. We will therefore summarize and characterize three of the studies that were reviewed by Nave et al. (2000). These studies illustrate the range of possibilities for applying the principles of experimental and quasi-experimental design.

Tennessee Class Size Study (1985–1989)

This experiment studied the effects of reduced class size on student achievement in kindergarten and grades 1, 2, and 3 in about eighty diverse public elementary schools throughout Tennessee (see also Mosteller, 1995). The initial study assessed student achievement in math and reading annually for four years. A follow-up study continued to assess the achievement of each group for additional years. We have combined both studies in an overall characterization of the employed experimental design. At the risk of oversimplifying the actual study design, we have represented it as follows:

R	X_A	O_1	O_2	O_3	O_4	. . .	O_k
R	X_B	O_1	O_2	O_3	O_4	. . .	O_k
R		O_1	O_2	O_3	O_4	. . .	O_k

The population of interest included all the kindergarten and first-, second-, and third-grade students and teachers in Tennessee. The study groups were nonrandom samples of 6,400 kindergarten students and 300 teachers. The involved schools, all of which had accepted invitations to participate, were from throughout the state; they were from inner-city, urban, suburban, and rural areas. Both the students and the teachers were randomly divided into and kept in three subgroups for four years. Treatment group 1 students and teachers (X_A) were in small classes of 13 to 17 students. Treatment group 2 students and teachers (X_B) were

in regular-size classes of 22 to 25 students and also had a teacher's aide. The control group students and teachers were in regular-size classes of 22 to 25 students and had no teacher's aide. Students' reading and arithmetic achievement scores were monitored across the four years of the study and subsequently across additional years. Comparisons between the groups showed that small class size had a positive and statistically significant effect lasting at least through the eighth grade. As a consequence, the Tennessee legislature allocated billions of dollars to lower class size in grades K through 3 in seventeen impoverished school districts. Subsequent investigation in these seventeen districts showed marked improvement in both math and reading test scores.

High-Scope Perry Preschool Study (1962–1965 and Beyond)

This experiment studied the effects of a rigorous preschool program for 123 disadvantaged children who were considered to be at risk of failure in school and later life (see also Schweinhart, Barnes, & Weikart, 1993). The treatment was delivered from 1962 through 1965, and treatment and control group participants were followed for more than thirty years. Hypothesized benefits of preschool experience included development of cognitive skills, success in school, graduation from high school, employment, economic self-sufficiency, positive family relationships, social responsibility, and staying out of jail. The design of this study could be configured approximately as follows:

Wave 0	R	X	O_1	O_2	O_3	O_4	O_5	O_6	. . .	O_k
	R		O_1	O_2	O_3	O_4	O_5	O_6	. . .	O_k
Wave 1	R		X	O_1	O_2	O_3	O_4	O_5	. . .	O_k
	R			O_1	O_2	O_3	O_4	O_5	. . .	O_k
Wave 2	R			X	O_1	O_2	O_3	O_4	. . .	O_k
	R				O_1	O_2	O_3	O_4	. . .	O_k
Wave 3	R				X	O_1	O_2	O_3	. . .	O_k
	R					O_1	O_2	O_3	. . .	O_k
Wave 4	R					X	O_1	O_2	. . .	O_k
	R						O_1	O_2	. . .	O_k

The study had five waves (cohorts) of participants, waves 0, 1, 2, 3, and 4 here, all from Ypsilanti, Michigan—two waves in the first year of the study and one in each of the three subsequent years. Each wave was randomly divided into a treatment group and a control group. The former received daily classroom sessions and home visits, whereas the latter received no preschool service. During the first year, the wave 0 group of four-year-olds participated in the preschool experience, and the wave 1 group of three-year-olds began their two-year stint of preschool. The treatment groups in waves 2, 3, and 4 were three-year-olds who all received two years of preschool experience. As feasible, all experimental and control participants were followed up with for more than thirty years. The evaluators applied a wide range of educational,

social, and economic measures to assess the program's success. Based on these measures, this program yielded many positive short-range and long-range benefits.

Career Academies Study (1992–2003)

This ten-year study was conducted to investigate the effects of career academies on high school students' academic achievement, progress toward graduation, and preparation for postsecondary education and employment (see also Kemple with Scott-Clayton, 2004). The design of this study can be characterized as follows:

Block A	R	X	O_1	O_2	O_3	O_4	. . .	O_k
	R		O_1	O_2	O_3	O_4	. . .	O_k
Block B	R	X	O_1	O_2	O_3	O_4	. . .	O_k
	R		O_1	O_2	O_3	O_4	. . .	O_k

Over a three-year period, about 1,700 volunteer eighth- and ninth-grade students across nine schools were randomly divided into a 959-member career academies group and an 805-member traditional high school (control) group. The experimental and control groups were blocked using high-risk (block A) and low-risk (block B) students as the randomized blocks. Student outcomes were monitored over a ten-year period using school records and other data on attendance, achievement, course-taking patterns, progress through high school, and post–high school performance. At the end of twelfth grade, high-risk students in the career academies group exceeded their control group counterparts in reduced dropout rates, improved attendance, increased participation in academic courses, and progress toward graduating on time. Of 490 students who completed twelfth-grade math and reading tests, however, no significant differences were found between the treatment and control students. Also, when all students' performance was analyzed, there were only minor advantages for the experimental group. This application of experimental design illustrates the increased precision that one can obtain by employing blocking variables for randomized experimental and control groups to look at effects on subgroups, such as males and females, minority and nonminority groups, and low- and high-risk groups (Raudenbush, Martinez, & Spybrook, 2007).

Guidelines for Designing Experiments

As Boruch (2003) noted, experimental design is very much a preordinate, or prospective, approach to evaluation. He advised evaluators to specify basic elements of a contemplated randomized field trial in advance. In the following paragraphs we summarize his advice as well as advice from Suchman (1967) and Shadish et al. (2002), and we also add some of our own.

Deciding Whether to Proceed with an Experiment

There are many circumstances warranting evaluation in which a randomized experiment would be inappropriate, not feasible, or not influential. Before proceeding to conduct an experiment, an evaluator should ascertain the feasibility of the contemplated study (with all its possible

social, legal, and ethical issues). In planning for the study, the evaluator should conduct an evaluability assessment—an essential process of determining whether the study would meet the following requirements:

- There is a well-defined treatment that can be implemented with fidelity.
- There is confusion about and/or a clear need for information on the treatment's effectiveness.
- There is (or can be obtained) sufficient clarity and consensus on the program's objectives and the values it should serve to warrant selection of credible and defensible outcome measures.
- The findings of the projected experimental study would address the study's purpose and effectively address stakeholders' questions.
- There are sufficient subjects to adequately power the study (though there are instances in which underpowered studies are warranted).
- There is assurance that study subjects will consent to being assigned randomly to experimental and control conditions.
- There is capacity, willingness, and agreement on the part of needed institutions to fully and faithfully play their respective roles in carrying out a randomized, comparative experiment.
- There is a strong, documented commitment from authority figures to maintain conditions necessary to faithfully implement the treatment and carry out the study over its full course.
- Program staff members understand and accept that the evaluation will not provide them with formative evaluation for continual program improvement.
- Program staff members are on record as agreeing to comply with randomization requirements and retain and hold the program treatment constant during the course of the experiment.
- Needed approvals from relevant institutional review boards and government organizations are in place to ensure that the needed information can be obtained.
- Those who are expected to participate in the treatment conditions and provide the needed information have given their written commitment to doing so.
- Any necessary guarantees of anonymity or confidentiality can be fulfilled.
- The participating evaluation staff possesses the technical expertise and availability to competently conduct all aspects of the study.
- There is reasonable written assurance that the resources required to carry out the full experiment will be provided.
- There is ample evidence that audiences for the evaluation report trust and have confidence in the evaluation team's integrity and competence.
- Overall, the contemplated experimental study would be feasible to conduct, be completely ethical, yield valid information on the program's effectiveness, and produce findings that would be used.

If any of the preceding conditions have not been met or cannot be met, the evaluator should not proceed with an experimental study and possibly should not conduct any other kind of study. Before declining the evaluation assignment, it would, however, be useful to consider whether some other evaluation approach would be feasible and credible and produce valuable information on the subject program. It is possible that persons' unwillingness to meet certain of these conditions would be mitigated if the evaluation were not constrained by requirements for randomization and experimental controls. For example, if program staff members knew they would get continual feedback to help improve the program and that using the feedback for program improvement was acceptable, they might be more forthcoming in supplying requested information and cooperating in other ways.

Values, Theory of Change, and Success Variables

As Suchman (1967) stressed, before proceeding with an experiment, an evaluator should determine and examine the values being sought through the program and the theory of how desired changes are to be obtained. These determinations provide the basis for identifying and assessing the needs of intended beneficiaries; defining the desired treatment; clarifying and assessing the ultimate, intermediate, and immediate program goals; determining which intended outcomes are targeted to which parts of the intended beneficiaries; and selecting appropriate outcome measures. It is almost a given that different stakeholders will have different, even contradictory ideas about what the program should do and not do and what it should achieve and not achieve (that is, conflicting values). Also, different intended beneficiaries might have different needs to be served by the program. Some persons are likely to worry that certain of the program's outcomes will be undesirable or harmful to particular program participants or other interested parties.

Before agreeing to conduct an evaluation, the evaluator is advised to meet with representatives of the full range of stakeholders, especially those who might be harmed by implementation of the subject program. When there are sharp value conflicts relating to the program, the evaluator should consider engaging stakeholders in a values clarification and consensus-building process. The evaluator also should judge whether the determined program goals are ethical and worth achieving.

If clear and defensible program goals have been identified, the evaluator can proceed to design a study based on the program's underlying values. As part of this process, the evaluator should seek to understand the client group's theory of how the program is expected to bring about the desired changes (also see Coryn, Noakes, et al., 2011; Funnel & Rogers, 2011). As Suchman (1967) recommended, the evaluator should define the independent and dependent variables, state the hypotheses to be tested, diagnose the social or technical problem or problems underlying the desire for change, clarify the desired changes and whether different changes are being sought for different subsets of beneficiaries, clarify the means of effecting the changes, and define the signs of actual change. If serious value conflicts cannot be resolved; if the stakeholders are unable to clarify the value system underlying the program; or if the program treatment, program goals, or both are unethical, the evaluator should decline to design and conduct the experiment.

Key Evaluation Questions

Having structured a randomized experiment, the evaluator is advised to exploit its power for addressing a wide range of relevant questions. He or she should consider addressing questions concerning the program's effects, such as the following:

- What were the quantity and quality of the effort to carry out the program?
- What were the program's intended and unintended outcomes?
- To what extent were program outcomes long range as well as immediate?
- To what extent did the program target and effectively address the needs of the intended beneficiaries?
- What were the program's differential effects on different parts of the experimental and control groups?
- What were the reasons for observed outcomes?
- How could program administrators make the program more effective or more efficient?
- Was the investment in the program justified by its outcomes?
- To what extent did program outcomes lead to changes in policies and practices inside and outside the program's geographical area?
- What are the program's implications for further research in the program area?

Populations, Units of Randomization, and Statistical Power

In many studies, it makes sense to define the population of interest and randomly select the members of the study sample from this group. A random sample taken from a predefined population of interest greatly enhances the possibility of extrapolating study findings to the larger population (Koleci, Coryn, Hobson, & Keci, 2011). In any long-range study, however, the population of interest at the study's conclusion could be substantially different from the original population. Such forces as gentrification can greatly change the composition and characteristics of a population in a particular area. Also, investigators rarely have sufficient control over sample selection to draw a random sample from a defined population. Usually it is necessary to deal with volunteers or other intact, nonrepresentative groups.

In such cases, about the best an evaluator can do is to randomly assign the members of the convenience group to experimental and control conditions. Random assignment is a powerful technique for ensuring equivalence of the comparison groups. The randomly assigned experimental and control units may be individual persons or intact groups. If the individuals are members of groups, such as classrooms, then their interactions with one another make them nonindependent units. In such cases, the need to meet the independence requirements of statistical analysis dictates that the group should be employed as the unit of analysis. However, this well-established statistical principle can be problematic. If using groups reduces the number of experimental units to only a few, then the power of the analysis to detect treatment

differences is low. Violating the assumption of independence by using individuals within groups as the experimental units may add statistical power. Sometimes it can be instructive to perform analyses on both individuals and groups and compare the results. Boruch (2003) advised that

> the main rules of thumb in assuring statistical power are (a) do a statistical power analysis, (b) match on everything possible prior to randomization and then randomize, (c) get as many units as possible, and (d) collect covariate/background data and time series data to increase the trial's power. (p. 113)

Interventions

The heart of any experimental study is one or more well-defined, powerful, and well-implemented treatments whose effects are of practical and often theoretical interest. The evaluator must proceed only on the basis that there is a definite treatment, that it will be faithfully implemented, that it will be held constant, that it will be applied only to the experimental group, and that its implementation can be monitored and verified. On this last point, it is essential that feedback from monitoring the treatment not become part of the treatment. If feasible, it is desirable that treatments be monitored, stabilized, and verified before starting an experiment. In advance of starting an experimental design evaluation, it is prudent to provide program staff with program guidelines and pertinent training. In general, experimental design evaluators should not waste time, resources, and effort in evaluating weak or phantom treatments.

Random Assignment

As noted repeatedly in this chapter, random assignment of subjects to treatments is a core feature of the experimental design approach. Random assignment is needed to ensure that every experimental unit has an independent and equal chance of being assigned to the experimental or control group. It should be done in an unbiased manner and close to the time when the treatment group will be involved in the program. The sample from which subjects will be randomly assigned should be large enough to ensure equivalence between the comparison groups. Boruch (2003) advised that "if it is possible to match or block, prior to randomization, this ought to be done" (p. 115). Prior to random assignment, the evaluator should obtain assurance that all units in the study sample are committed to participating or will be required to participate faithfully in the group to which they are randomly assigned, for the full duration of their intended involvement.

In terms of mechanics, random assignment preferably is done through the use of an appropriate software package or a random numbers table. The randomization process should be managed by a person with no vested interest in the program. Boruch (2003) advised against such loose approaches as coin tosses and pulling numbers from a hat because they are prone to bias. (See also Shadish et al. [2002] for suggestions in regard to random assignment of units to conditions.)

Observation, Measurement, and Theory

The modern concept of experimental design requires a comprehensive approach to data collection. The evaluator should collect both quantitative and qualitative information and should be careful not to collect unneeded information. The required information includes the background characteristics of each person or entity in the treatment and control groups, the program's context, relevant needs assessment information, intended and actual beneficiaries, field notes on how and how well the experimental treatment was implemented, the activities of persons or entities in the control group, a record of program costs, differential effects, intended and unintended outcomes, and responses to all agreed-on investigatory questions. In choosing measurement variables, it can help the evaluator to consider the elements of a relevant theory that denote how particular treatments are expected to produce desired outcomes. In assessing the implementation of a treatment, the evaluator should record and keep a count of instances of poor implementation and nonimplementation plus qualitative information on the reasons for noncompliance. Such data should be recorded in a detailed technical appendix to the evaluation report. In certain instances, measures of implementation fidelity can be used to investigate whether treatment integrity moderates an observed treatment effect.

Management

Evaluations based on experimental design principles require sustained, competent, and effective management. The study manager must secure the required funding; establish and maintain mutual trust among the involved parties; secure needed cooperation from various individuals, groups, and institutions; appoint and involve advisory panels; involve other interest groups; schedule the needed work; delegate authority and responsibility for different parts of the study; recruit, train, and supervise staff; anticipate and cope with political threats to the study; maintain control of the implementation of the experimental design; take necessary steps to ensure that experimental and control groups are kept separate; maintain communication with all interested parties; foster positive public relations for the study; host visiting dignitaries, researchers, content experts, program stakeholders, and members of the media; maintain fiscal accountability; meet all legal requirements; write and deliver necessary management reports; report on progress to the evaluation's sponsor, client, and professional reference groups; subject the study's plans, process, and reports to independent metaevaluation; foster dissemination of study findings; and troubleshoot and solve problems as needed. Clearly there is much more to conducting a sound, experimentally oriented evaluation than just carrying out the technical work. The evaluation management task is essentially the same for all types of evaluation.

In spite of the evaluator's best efforts to implement the needed controls, the conditions required to conduct a successful experiment might break down at any point during the study process. Therefore, it is prudent for the evaluator to prepare contingency plans in case this happens. For example, the evaluator might be prepared, if necessary, to convert the randomized controlled experiment to a quasi-experiment or even a case study.

Analysis

Data from experiments should be thoroughly and systematically analyzed. Appropriate analyses (to establish statistical conclusion validity) are necessary to determine whether the experiment was carried out as planned, whether the treatment condition of interest produced better results than a control condition or alternative treatment, whether observed differences are practically important, whether the program was more effective for some subgroups of the treatment sample than for others, and whether the findings may be extrapolated to other settings. Following are general classes of analysis needed to address this range of issues.

Design Implementation Analysis

This analysis is necessary to determine the extent to which the treatment (or treatments) was implemented as intended and applied to the intended beneficiaries, and to confirm that experimental and control (or alternative treatment) groups did not differ in important ways at the experiment's outset and did not become contaminated by each other along the way. The evaluator should compile and analyze field notes on how and how well the treatment was carried out; the extent to which experimental and control conditions were applied consistently and exclusively to the intended groups; any major deviations from the intended treatment process; and, in general, the fidelity of the implementation of the treatment. Most of these analyses will be qualitative in nature. The evaluator should assess the extent to which the comparison groups were equivalent by creating tables comparing the groups at the experiment's outset on such variables as socioeconomic status, ethnic origin, age, gender, school years completed, aptitude test scores, and grade point average.

Core Analysis

What Boruch (2003) has referred to as "core analysis" addresses the central question in any experiment: Is the observed difference between the comparison groups statistically significant? Such analysis estimates the magnitude of difference between the assessed outcomes for the groups and, based on the theory of randomization, gives a statistical statement of the level of confidence that the difference is not due to chance variation. Determinations of statistical significance follow the rules of *t*-tests, analysis of variance, analysis of covariance, and other relevant tests. Based on the theory of randomization, these tests indicate the number of times (such as less than one or less than five) out of one hundred that the observed difference would be expected based on chance variation. Any observed difference beyond the chosen 0.05 or 0.01 significance level is judged to be statistically significant. As Boruch (2003) emphasized, tests of statistical significance must analyze outcome measures for the comparison groups as they were originally randomly assigned. This is required to adhere to the statistical theory and logic associated with comparing equivalent groups. If by some circumstance some members of an experimental group switched to the control group (or vice versa) during the course of the experiment, it is not permissible to switch them from their original assigned group for purposes of analysis. Data on them must be included with the data for the group to which they were originally assigned. This is essential to preserve the validity of the statistical analysis and

also ensures conservatism in claiming a statistically significant difference. Good contemporary statistical practice, however, dictates that treatment and control group differences be reported in the form of an appropriate effect size metric (for example, standardized mean difference, odds ratio, or risk ratio) and its corresponding 95 percent or 90 percent confidence interval (that is, the level of precision). This would be done in addition to, or instead of, conducting simple tests of statistical significance (Kline, 2004, 2008). (Effect sizes and confidence intervals are discussed at length in Chapter 23.)

Practical Significance

Because policymakers or program stakeholders may not always consider a statistically significant difference important, the investigator often will conduct further analysis to help the interested parties reach a judgment of whether the observed difference has practical significance (again, often through reporting standardized effect sizes and their corresponding 90 percent or 95 percent confidence interval). More typically, judgments of practical significance will be secured through qualitative means. For example, before findings are obtained, the investigator might engage focus groups to discuss and reach agreement on what level of difference would be sufficiently important to adopt an innovation being compared with current practice. Following a determination of statistically significant findings, the evaluator might reconvene the focus group, confront the participants with their previous conclusions on practical significance and the findings on statistical significance, and engage them to consider whether the statistically significant differences obtained are sufficiently important to warrant adoption of the experimental approach.

As another approach, the investigator might conduct a cost analysis (see Chapter 6) to help decision makers consider whether the possible improved performance available from adopting a new practice is worth the projected cost of adopting the approach (which could be more or less than the cost of current practice). For example, a school board might not be willing to bear a substantial cost associated with reduced class size if the statistically significant improvement in test scores is only two or three points on a standardized test.

A Posteriori or Post Hoc Tests

Often the core analysis will yield a statistically significant interaction between a treatment variable and blocking variables, such as gender or socioeconomic level. In general, interactions frequently indicate that although treatment differences might not be statistically significant overall, there could be significant differences between treatment and control condition results for certain subgroups. For example, reduced class size might not show statistically significant improvement for a total group of students but could reveal significant differences between treatment and nontreatment students from impoverished homes. Probing of such observed subgroup differences through a posteriori or post hoc tests may not yield findings that are supportable from the underlying theory of random assignment, but it can yield hypotheses for further investigation and policy-level deliberation.

Literature Reviews

Still another form of analysis is to conduct a literature review. One would conduct a systematic search of the relevant literature to identify studies that were similar to the subject experiment. One's experiment and the relevant similar studies found in the literature would then be compared. They would be contrasted on design, study samples, procedures, and findings. The central question would be whether the current experiment has produced findings that coincide with those of quite similar studies. Differences in findings would be discussed in regard to whether they might represent instability in one's findings or possibly be attributable to other factors concerning when the studies were conducted and whether, for example, they differed substantially in study samples, treatments, social context, and criterion measures. The results of such an analysis would be useful in assessing the firmness of one's findings and possibly their generalizability (see also "Approach 14: Meta-Analysis" in Chapter 6).

Reporting

In accordance with the preordinate nature of randomized experiments, agreements on reporting evaluation findings should be determined at the study's outset and recorded in the study's contract. Key reporting agreements, according to the Joint Committee on Standards for Educational Evaluation's *Program Evaluation Standards* (1994), include the following:

- Rightful report recipients. Often these are the client that commissioned the evaluation, those legally responsible for the subject program, those who in some way helped fund the program, those who contributed a substantial amount of information to the evaluation, and other stakeholders who are quoted in or will be affected by the report.
- Use of reports, and report formats that are appropriate for different intended users. Reports may include a main report of the study's questions, experimental subjects, design, data collection procedures, and findings; an extensive technical report containing instruments, procedures, a description of study subjects, analyses, and evaluator qualifications; a report of an independent audit or metaevaluation; a journal article; or a computer-assisted presentation of main points from the study.
- An appropriate schedule of reporting keyed to both the times when different parts of the information will be available for release and the needs of the intended users.
- Specification of authority over and responsibility for finalizing and disseminating evaluation findings. The evaluator must retain basic authority over preparing findings and editing reports. Moreover, the evaluation contract should ensure that the evaluator's final report will be given to an agreed-on audience. If different reports are to be given to different parts of the audience, this should be specified in the contract along with the proviso that the evaluator will control content and editing. Typically, the evaluator should obtain firm written agreement that the client will assist with getting the appropriate reports to the appropriate subsets of intended users.

Metaevaluation

As with all other approaches to program evaluation, experimental design studies should be subjected to formative and summative metaevaluations. Preferably the summative metaevaluation is conducted by an independent methodologist or a team of methodologists, whereas it is probably more realistic for the evaluator to conduct and document the formative metaevaluation. Formative metaevaluation involves monitoring and providing feedback on the extent to which the evaluation study is adhering to the original design. Summative metaevaluation entails assessing the extent to which the evaluation has met the requirements of utility, feasibility, propriety, accuracy, and evaluation accountability as defined in Chapter 3.

Given the centrality of internal validity, metaevaluations of experimental design evaluations should, at a minimum, consider the plausibility of internal validity threats, as detailed earlier in the chapter (Shadish et al., 2002).

Quasi-Experimental Designs

Because so often it is not feasible to implement a true experimental design, evaluators and other researchers often are forced to rely on quasi-experimental designs (Campbell & Stanley, 1963; T. D. Cook & Campbell, 1979; Shadish et al., 2002). In certain circumstances, these may be quite feasible because they do not involve random assignment of participants. They do, however, require that the treatment be well defined and implemented and that the control group, if there is one, be kept separate from the experimental group. In addition to some of the quasi-experimental designs discussed earlier in this chapter, two particular quasi-experimental designs are powerful alternatives to experimental designs: regression discontinuity designs and interrupted time-series designs. Although both regression discontinuity and interrupted time-series designs are viable options in many circumstances, neither design is used with much frequency in program evaluation. T. D. Cook, Shadish, and Wong (2008) and Shadish, Galindo, Wong, Steiner, and Cook (2011) have recently begun exploring the conditions under which regression discontinuity designs produce causal estimates comparable to those derived from experiments.

Regression Discontinuity Designs

Regression discontinuity designs require control over assignment of participants to one or more treatment and control conditions through pretesting. Unlike experimental designs that require random assignment to conditions, regression discontinuity designs assign units to conditions on the basis of a cut score on an assignment variable, often based on treatment need or merit. Therefore, regression discontinuity designs are particularly useful for alleviating some of the objections frequently expressed when random assignment is used, sometimes depriving individuals of needed treatments or services. Another advantage of regression discontinuity designs, as with randomized experiments, is that the method of assignment is completely known and perfectly measured, which is not true of any other quasi-experimental design, thus eliminating or allowing modeling of selection bias (see Shadish et al. [2002] and Trochim [1984] for detailed discussions). It also should be noted, however, that regression discontinuity

designs require larger numbers of subjects for adequate statistical power than do experimental designs (Trochim & Cappelleri, 1992).

The basic regression discontinuity design can be represented as follows:

O_A	C	X	O_1
O_A	C		O_1

Here, O_A is a preassignment measure of an assignment variable and C is a cut score on the assignment variable. If j is a cut score on the assignment variable (notated as O_A) then any score greater than or equal to j indicates assignment to one condition and any score less than j indicates assignment to the other condition. However, "the assignment variable must have at least ordinal characteristics, that is, be monotonically increasing; true nominal variables such as ethnicity are specifically excluded" (Shadish et al. 2002, p. 209). Often, the assignment variable is a pretest of a posttest outcome variable, but it is not required that the assignment variable be correlated with the outcome variable. The cut score can be placed at any point on the assignment variable, but cut scores set near or at the mean of the assignment variable are typically more powerful.

Shown in Figure 11.2 is an illustration of a hypothetical regression discontinuity study in which participants scoring at or above the cut score were assigned to receive treatment and those below the cut score were not. In the figure the vertical line at the cut score on the assignment variable partitions the groups into treatment and control conditions. The vertical

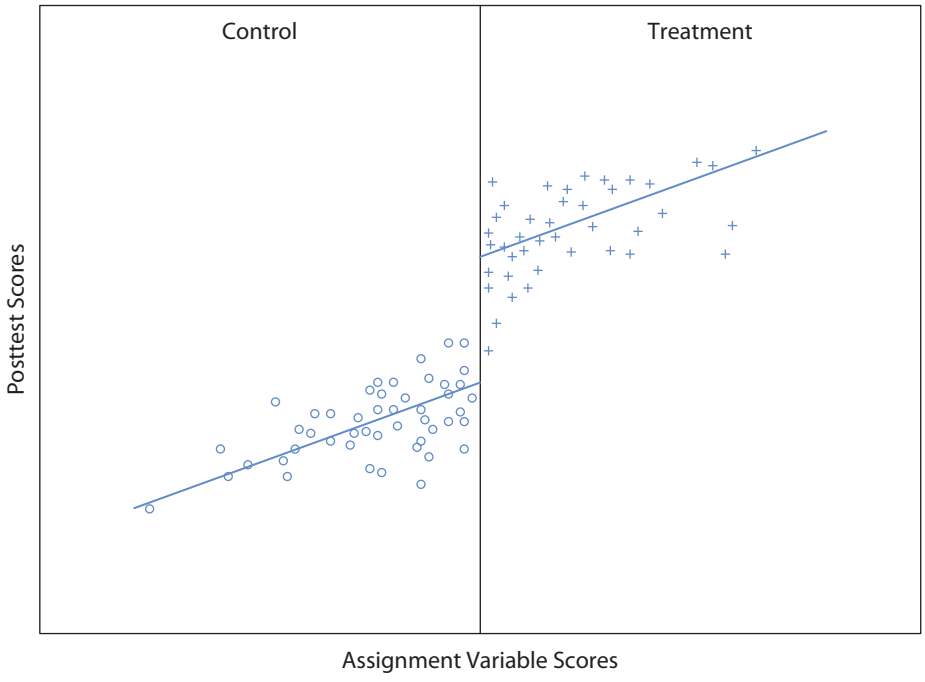

Figure 11.2 Hypothetical Regression Discontinuity Study of an Effective Treatment

displacement (that is, discontinuity) of scores at the cut point indicates a positive treatment effect—here a change in intercept, although other functional forms (such as a change in slope) are possible. The displacement should occur at exactly the point on the assignment variable at which the cut score defines the treatment contrast.

Figure 11.3 shows the results of a hypothetical regression discontinuity study in which there was no differential effect on students' performance on a test of reading comprehension following an eight-week after-school reading program. The hypothetical example in Figure 11.4 shows a small, positive effect on word processing speed for the subjects who received specialized word processing instruction compared to those who did not. The hypothetical example from health care, shown in Figure 11.5, shows a significant reduction (represented by a change in slope, that is, the steepness of the regression line, rather than a change in intercept) in ferritin iron measures for subjects receiving six months of weekly phlebotomies compared to those who received no such treatments. As a final example, Figure 11.6 shows a positive effect on mathematics test scores for students who received a special fifteen-week mathematics improvement program.

The statistical analysis of regression discontinuity designs depends on the shape of the regression surfaces between the assignment variable and the outcome variable in the two groups. Often, the functional form is treated as linear (that is, a straight line). However, if the functional form is of a different shape (for example, curvilinear), a linear model will produce biased estimates of the treatment effect (Reichardt, 2005).

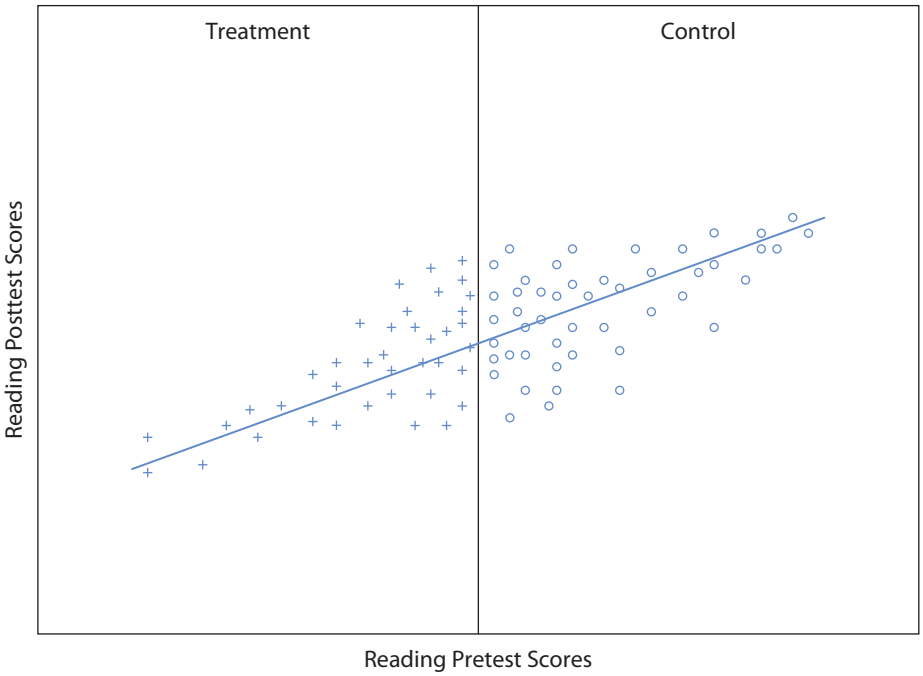

Figure 11.3 Hypothetical Regression Discontinuity Study Showing No Effect on Reading Comprehension Test Scores for Students Who Received an Eight-Week After-School Reading Program

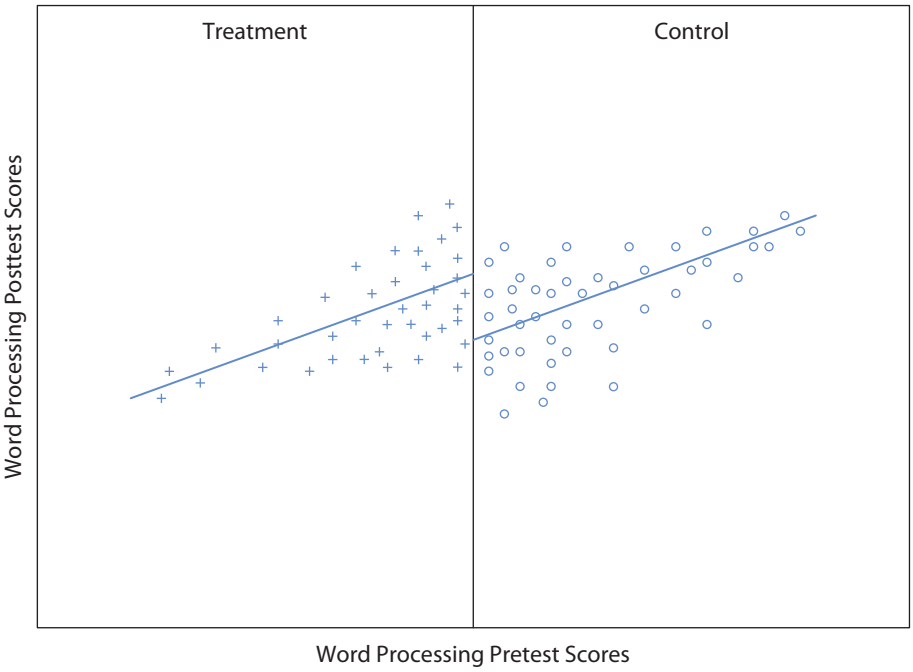

Figure 11.4 Hypothetical Regression Discontinuity Study Showing a Positive Effect on Word Processing Speed for Students Who Received Four Weeks of Word Processing Instruction

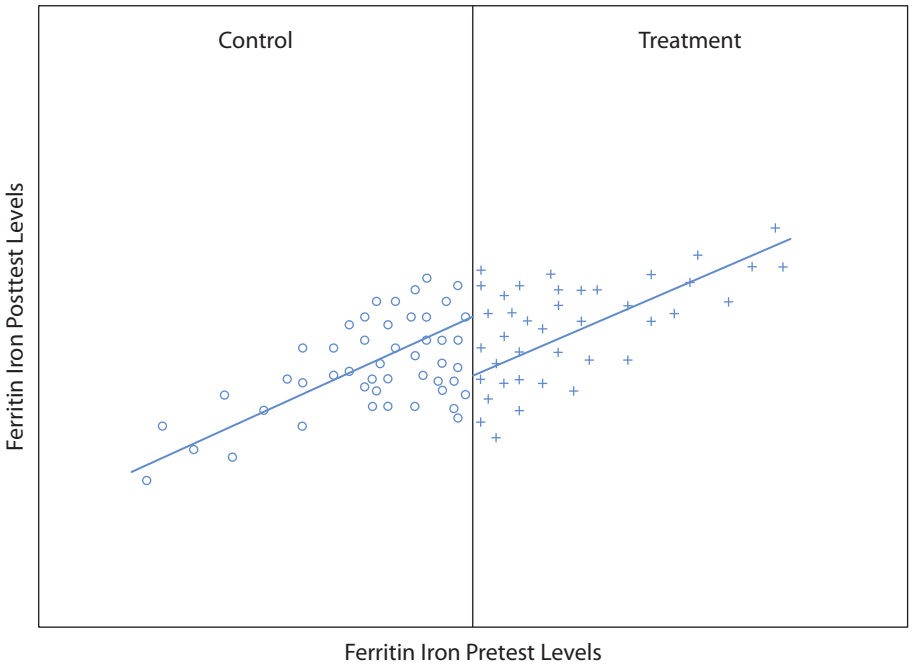

Figure 11.5 Hypothetical Regression Discontinuity Study Showing a Positive Effect on Lowering Ferritin Iron Levels for Patients Who Received Weekly Phlebotomies over a Six-Month Period

Figure 11.6 Hypothetical Regression Discontinuity Study Showing a Positive Effect on Mathematics Test Scores for Students Who Received a Fifteen-Week Mathematics Improvement Program

Interrupted Time-Series Designs

Interrupted time-series designs are a class of designs in which there is a large series of observations made on the same variable over time. Although not as common, short interrupted time-series studies are often a feasible alternative to randomized experiments. In these types of designs,

> observations can be made on the same units, as in cases in which medical or psychiatric symptoms in one individual are repeatedly observed . . . [or] . . . different but similar units, as in cases of traffic fatalities displayed for a particular state over many years, during which time the baseline population is constantly changing. (Shadish et al., 2002, p. 172)

The central feature of interrupted time-series designs is knowing exactly when a treatment or program was introduced. If a treatment or program had an effect, then observations occurring after the treatment or program will have a different slope or level (that is, the "interrupt") than those occurring prior to the introduction of the treatment or program.

The basic one-group interrupted time-series design with multiple pretest and posttest observations can be expressed as follows:

O_1	O_2	O_3	O_4	O_5	X	O_6	O_7	O_8	O_9	O_{10}

This design can be extended to include a nonequivalent or equivalent control group, which would be diagrammed as follows:

O_1	O_2	O_3	O_4	O_5	X	O_6	O_7	O_8	O_9	O_{10}
O_1	O_2	O_3	O_4	O_5		O_6	O_7	O_8	O_9	O_{10}

Other design elements to strengthen interrupted time-series designs include removed treatments, nonequivalent dependent variables, multiple replications, and switching replications, all of which can be used to reduce potential internal validity threats. Although interrupted time-series studies are sometimes good alternatives to randomized experiments, a number of problems frequently arise in conducting these types of investigations. With respect to using such designs for program evaluation, two problems are particularly salient (Shadish et al., 2002):

- Many treatments are implemented slowly and diffuse through a population, so that the treatment is better modeled as a gradually diffusing process rather than as occurring all at once.
- Many effects occur with unpredictable time delays that may differ among populations and over time.

Shown in Figure 11.7 is a plot of incidence rates of sexually transmitted diseases (STDs) in two high schools over a period of more than twenty years. As shown in the figure, the treatment (a new sex education curriculum), indicated by the horizontal line, was introduced in one school (that is, the treatment school) midway through the time series, but in not the other.

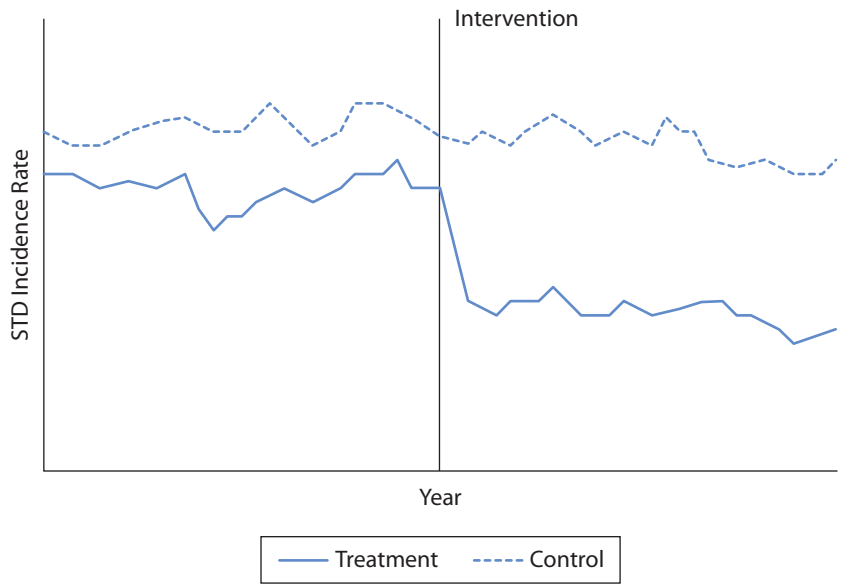

Figure 11.7 Effect of a New Sex Education Curriculum on STD Rates

Here, the addition of the control school greatly reduces, but does not completely eliminate, many alternative explanations for the dramatic reduction in STDs in the treatment school.

Time-series data are typically autocorrelated (that is, they are serially dependent, whereby the value of one observation is usually related to the value of previous observations that may be one, two, three, or more time lags away), and a substantial number of observations are often required to correctly model the autocorrelation. Classical statistical procedures, such as ordinary least squares regression, assume that observations are independent rather than correlated. Autocorrelation does not bias estimates of treatment effects, but it tends to lead to an underestimation of standard errors, which makes statistical significance tests too liberal and confidence intervals too narrow (Reichardt, 2005). Because autocorrelation concerns only the stochastic component (the effect of a random event occurring at the time of the observation and the effect of previous random events), it does not bias deterministic estimates, such as the mean. It does, however, bias estimates of the error variance and, therefore, all conventional tests of statistical significance. In the most common situation, whereby observations are positively autocorrelated, the error term is increased and creates a liberal bias when ordinary hypothesis testing procedures are used. That is, interventions may be found to be statistically significant when no real effect exists. The effects of autocorrelation can be accounted for using such statistical procedures as autoregressive integrated moving average modeling, among others, for statistical analysis of interrupted time-series designs. Appropriate statistical modeling and analysis for both regression discontinuity and interrupted time-series designs are technically demanding, and interested readers are referred to Murnane and Willett (2011) and Cryer and Chan (2010), respectively.

Summary

Randomized controlled experimental design was quite prominent in program evaluations during the late 1960s and early 1970s because the U.S. government required its use to evaluate federally funded innovations in education and other social services. Experimental design's consistent, widespread failure to produce useful information, especially in education, caused it to lose favor and be replaced by a host of new program evaluation approaches, especially from the mid-1970s through the 1990s. Ironically, the federal government in this century has once more mandated the use of randomized experiments to evaluate federally funded educational programs and has been holding federal evaluation funds hostage to this requirement. We see this as a serious error in judgment. Although there have been some successful experimental design evaluations of educational and social programs, sound experiments that have produced useful information about such programs have been rare. Currently many sound evaluation models and approaches capture the nuances of a program and provide useful formative feedback in ways that the experimental design approach never can. Nevertheless, pioneers using experimentation to evaluate programs, including Boruch (2003); Campbell and Stanley (1963); T. D. Cook and Campbell (1979); Cronbach and Snow (1969); Shadish et al. (2002); and Suchman (1967), have played a valuable part in the program evaluation field and recently have been advancing new methods for using experimental and quasi-experimental designs.

In our effort to be fair and balanced, we devoted most of this chapter to describing the foundations of experimental design and presenting guidelines for deciding when and how to apply the approach. It is quite possible, of course, for an experimental design to be used as part of a wider evaluation, and we acknowledge that the prevalent use and success of experiments in such fields as medicine and agricultural sciences have justifiably led experimental design to gain considerable credibility. It is especially when the units receiving treatments are in social groups, such as schools, that randomized experimental design has severe limitations. We stress that randomized controlled experimentation must never be construed as the single best methodology in most program evaluation situations. We agree with Scriven (2005a) that the appropriate gold standard of cause-and-effect studies is not conclusions from a randomized controlled experiment, but conclusions beyond reasonable doubt, whatever evaluation approach is applied. In field situations in which true experiments are appropriate and feasible, the practices of randomization and control of treatments can produce sound and useful information on program effectiveness; however, such circumstances are rare. Nevertheless, this chapter has conveyed valuable information based on the works of such authors as Suchman, Boruch, Shadish, Campbell, and Cook that evaluators should consider when client needs and program conditions make feasible the conduct of rigorous experimental studies. Clearly, the described cases in this chapter make clear that there are occasions when such experiments can make valuable contributions to understanding the effects of significant interventions.

REVIEW QUESTIONS

1. Identify and define the most basic methodological requirement for conducting an evaluation based on a randomized controlled experimental design; then list the more general requirements for conducting a sound experimental study.
2. Distinguish between studies that assess an intervention's effects and ones that assess the causes of observed outcomes; then discuss the applicability of randomized controlled experimental design to both types of studies.
3. List what you see as Suchman's most important contributions to current concepts and methods of program evaluation. In particular, what were his key contributions to the use of experimental designs in program evaluation?
4. What did Suchman list as the principles of evaluation research design?
5. Nave, Miech, and Mosteller as well as Boruch concluded that randomized controlled experiments are rare in education. Considering that the federal government repeatedly has mandated use of this approach, why do you think it is rarely applied successfully in education?
6. List and briefly describe at least three examples of successful and socially significant evaluations based on randomized controlled experimental design.

7. List and define at least six threats to the validity of an experimental design study's findings.
8. In a true experiment, what are the preferred methods of randomly assigning subjects to treatment and control groups? What reasons do you think Boruch would give in advising that a program director not do this assigning him- or herself and that assignment not be based on coin tosses or drawing names from a hat?
9. Define and give an example of each of the following types of analysis employed in experimental studies: design implementation analysis, core analysis, practical significance analysis, a posteriori and post hoc tests, and literature review analysis.

Group Exercises

Exercise 1

Suchman was one of the first to endeavor to incorporate ideas and knowledge about evaluation into the implementation of new programs, or changes in existing programs, in the public sector. Discuss this point about Suchman along these lines:

1. What are some of the main purposes Suchman saw for evaluation research?
2. Using his concept of value, how would he identify and define the criteria for researching and judging a program, and what would be his approach to reaching the final judgment of the program?

Exercise 2

Identify and summarize a program with which your group is familiar and that you see as potentially amenable to a randomized controlled experiment. Then make a list of this chapter's contemporary guidelines for designing experiments, and follow these steps:

1. For each guideline, write a "yes" or a "no" in regard to its acceptability for evaluating this program, and provide a brief explanation of your group's decision.
2. Present and justify your conclusion on whether a randomized controlled experiment is appropriate to evaluate the program and, if so, what type of experiment would be appropriate.
3. If you conclude that a randomized controlled experiment is not appropriate to evaluate this program, what other approach would your group be likely to use? Why?

Exercise 3

Suppose your group is writing a job description for the person in charge of managing an experiment to be conducted on the program you identified in your response to exercise 2. List the main responsibilities for managing the experiment.

Suggested Supplemental Readings

Bigman, S. K. (1961). Evaluating the effectiveness of religious programs. *Review of Religious Research*, *2*, 108–109.

Boruch, R. F. (2003). Randomized field trials in education. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 107–124). Norwell, MA: Kluwer.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on training* (pp. 171–246). Skokie, IL: Rand McNally.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Skokie, IL: Rand McNally.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Skokie, IL: Rand McNally.

Cook, T. D., Scriven, M., Coryn, C.L.S., & Evergreen, S.D.H. (2010). Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven. *American Journal of Evaluation*, *31*, 105–117.

Coryn, C.L.S., & Hobson, K. A. (2011). Using nonequivalent dependent variables to reduce internal validity threats in quasi-experiments: Rationale, history, and examples from practice. In S. Mathison (Ed.), *Really new directions in evaluation: Young evaluators' perspectives* (pp. 31–39). New Directions for Evaluation, no. 131. San Francisco, CA: Jossey-Bass.

Cronbach, L. J., & Snow, R. E. (1969). *Individual differences in learning ability as a function of instructional variables*. Redwood City, CA: Stanford University Press.

Fisher, R. A. (1951). *The design of experiments* (6th ed.). New York, NY: Hafner.

Fleck, A. C. (1961). Evaluation as a logical process. *Canadian Journal of Public Health*, *52*, 185–191.

Herzog, E. (1959). *Some guidelines for evaluative research*. Washington, DC: U.S. Department of Health, Education, and Welfare.

James, G. (1958). Research by local health departments—Problems, methods, results. *American Journal of Public Health*, *48*, 354–379.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Klineberg, O. (1955). The problem of evaluation. *International Social Science Bulletin*, *7*, 347–362.

Nave, B., Miech, E. J., & Mosteller, F. (2000). A rare design: The role of field trials in evaluating school practices. In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed., pp. 145–162). Norwell, MA: Kluwer.

Scriven, M. (2005). *Can we infer causation from cross-sectional data?* Washington, DC: National Academy of Sciences.

Scriven, M. (2005). Causation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 43–47). Thousand Oaks, CA: Sage.

Smith, M. F. (1989). *Evaluability assessment: A practical approach*. Norwell, MA: Kluwer.

Suchman, E. A. (1967). *Evaluative research: Principles and practice in public service and social action programs*. New York, NY: Russell Sage Foundation.

Trochim, W.M.K. (1984). *Research design for program evaluation: The regression discontinuity approach*. Thousand Oaks, CA: Sage.

CHAPTER 12

CASE STUDY EVALUATIONS

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **The case study approach's logic, information requirements, naturalistic methods, and types of reports**
- **Robert Stake's approach to case study evaluations**
- **Robert Yin's approach to case study evaluations**
- **Particular methods for gathering case study information**

This book places the case study approach to the investigation of a naturalistic phenomenon under the general category of quasi-evaluation, as described in the first edition of this book (Stufflebeam & Shinkfield, 2007), though not discussed in this edition due to space constraints. The dominant reason for this categorization is that some applications of the approach are focused more narrowly than needed to assess a program's merit and worth, whereas other applications are configured to fully assess a program's value. Accordingly, the case to be evaluated might be a total program, some component of a program (such as its annual budgeting process), or the situation and experiences of one or more individuals being served by a program. In this chapter we discuss the case study approach mainly as it applies to evaluating a total program. Depending on the circumstances, the case study approach may be the optimum way to examine and illuminate a total program.

Overview of the Chapter

This chapter is divided into four main parts. In the first we provide a general description of the case study approach. The subsequent part describes and discusses Stake's seminal contributions in applying the general case study approach to the realm of program evaluation. In the third part we look at Yin's adaptation of case study methodology for program evaluation. Subsequently, in the final part, we add our views on appropriate methods for gathering case study information.

Overview of the Case Study Approach

We begin this part of the chapter by looking at the most fundamental requirements of a case study, particularly its noninterventionist nature and the necessity of rich, in-depth description of the subject case. Then we discuss particular elements of a case study, including the types of information needed, methods, special considerations related to sampling, and reporting of findings.

Essence of a Case Study Program Evaluation

A case study evaluation's signature feature is an in-depth, noninterventionist examination of the case and issuance of a captivating, illuminative report. The evaluator closely observes and meticulously records the case in its natural setting. The evaluator studies, analyzes, and describes the case as fully as possible. He or she obtains and reviews pertinent documents, interviews principal parties involved in the case or in a position to share insights about the case, and possibly collects pertinent photographic evidence. He or she examines the case's context, goals or aspirations, plans, resources, unique features, importance, noteworthy actions or operations, achievements, disappointments, needs and problems, and other topics. Ultimately the evaluator prepares and issues an in-depth report on the case, with descriptive and judgmental information, perceptions held by different stakeholders and experts, and summary conclusions.

Noninterventionist Nature of Case Study Evaluations

It is important to note that evaluators do not control the program (or program component) in any fashion as they might, for instance, if they were applying an experimental design. This approach is not confined to tightly controlled, formalized collection and analysis of data.

Scope of Case Study Information

Case study investigators closely examine the context, including program participants' needs, inputs, operations, intended and unintended effects, and any other processes (with all their complexities) that are producing outcomes. Focus is placed on portrayal of events, testimonials, stored information, and individuals involved in program implementation and direction, so that stakeholders are given information for understanding the program and making needed improvements. This information necessarily will depict the multidimensional nature of the setting in which a program is in progress.

One important aspect of case studies is the appropriateness of informant selections, program locations and occasions, and materials for interviews and other modes of data collection. This consideration may be simplified if, for example, there is a single group of program stakeholders or a dominant group representing or participating in a program.

Sampling Issues

In many case study evaluations, the selection of data sources is difficult and problematic, and an evaluator cannot use probability sampling (where, for example, program participants are selected randomly to participate in interviews or surveys) methods to obtain information representative of a population of interest. There is no simple solution to this ever-present problem for case study evaluators.

To obviate the effects of nonprobability sampling, evaluators must assess the field of potentially useful respondents to ascertain the representativeness of the overall body of program participants (or decision makers or other stakeholders) and record this information as completely as possible for future reporting. If this is done well, divergence as well as convergence of opinions and approaches to a program will be captured, so that a holistic and properly representative view of a program and its environment is ultimately possible.

Clearly, a context evaluation, which includes an assessment of the needs of program participants and beneficiaries, would be a sound starting point. It would give credibility to decisions made later about selections of individuals, groups, and events for close involvement in a case study. Also, a cardinal principle of case study evaluations is to continue identifying and querying data sources until no further insight would be gained by gathering additional information. On this point, it is essential to note that a sufficient set of case study information often will include conflicting accounts, as might be expected, for example, if one were interviewing Democrats and Republicans about the merits of a proposed piece of legislation.

Case Study Methods

Using as many methods as necessary, case study evaluators view a program in its different (and possibly opposing) dimensions as part of presenting a general characterization of the case. As emphasis is placed on the ethnographic nature of the program, it is likely that qualitative techniques will be used, with experienced, professional judgment as an ever-present complement to the study. Case study evaluators watch, listen, and interview, and they follow up on trails of interest, doubts, and perplexities until they are able to present a full account of a program.

Reporting of Case Study Findings

Final reports are usually written for appropriate program decision makers, other stakeholders, and other interested parties. However, recommended courses of action will often be difficult to state with a high level of confidence. This is not to suggest that the case study approach leaves stakeholders languishing in decision-making limbo. Quite the opposite is true. A sound case study provides abundant information for decision making with a clear teasing out of the intricacies that abound in naturalistic settings. Case study evaluators may report findings only to help a broad audience understand a program and reach their own judgments. Or they may tailor a report to the needs of decision makers by including an array of possible solutions to problems and other ways to improve a program.

Case Study Research: The Views of Robert Stake

Stake introduced his 1995 book, *The Art of Case Study Research,* in this way:

> A case study is expected to catch the complexity of a single case. A leaf, even a single toothpick, has unique complexities—but rarely will we care to submit it to case study. We study a case when it itself is of very special interest. We look for the detail of interaction with its context. Case study is the study of the particularity and complexity of a single case, coming to understand its activity within important circumstances. In this book, I develop a view of case studies that draws from naturalistic, holistic, ethnographic, phenomenological, and biographic research methods. (p. xi)

He went on to say that there are many kinds of case studies, all with their place (Stake, 1995). For example, there are quantitative case studies based on measurement of descriptive variables, often used in medicine, and case studies constructed for instructional purposes, common in colleges of business and law. However, Stake's interest (at least as outlined in his recent writings), is disciplined, qualitative inquiry. "The qualitative researcher emphasizes episodes of nuances, the sequentiality of happenings in context, the wholeness of the individual" (Stake, 1995, p. xii). Stake contends that this approach is an effective way of studying educational programs generally, and one that is particularly useful for program evaluation. Although he personally organizes the case study around identified issues, he advocates that those wishing to pursue case studies be aware that certain other techniques could prove more satisfactory, depending on idiosyncratic style or prevailing circumstances, especially when the object to be studied is more a relationship or a phenomenon than an explicit case. In other words, a case study is defined not by a methodology, but by the choice of object to be studied. A case study evaluation's main purpose is to provide stakeholders and other audiences with an authoritative, in-depth, and well-documented interpretation of a program.

The Value of Drawing Conclusions About Single Cases as Opposed to Seeking Generalization

Concerning the choice of a name for qualitative studies of single cases, some have preferred the term *fieldwork* to *case study*. Stake's choice of the latter term for evaluation studies resides in the attention it draws to the question of what specifically can be learned about the evaluand: "That epistemological question is the driving question: What can be learned from the single case? I will emphasize designing the study to optimize the understanding of the case rather than generalization beyond" (Stake, 1994, p. 236).

The Issue of a Case Study's Duration

According to Stake (1994, 1995), a case study need not be bound by time. One study may take a few weeks of intensive fieldwork preceded by planning time and followed by close analysis of documentation and writing, entailing some months in all. Another may require even less

time—perhaps a week or so—to achieve its aim. And others run for years, depending on the number and magnitude of the issues under focus.

Whatever the duration of the study, the general conceptualization does not differ significantly. An important responsibility is to sharpen identification of the case, whatever it may be, and concentrate on it for as long as it takes to understand its complexities. Moreover, Stake (1988) averred that these complexities, such as multiple sponsorship of an innovative program, are subject to the program's context and dynamics, including its geographical boundary, cultural and social environment, and patterns of behavior, which are important in understanding the case.

Case Study Types

Stake (1994) identified three types of case studies. The *intrinsic case study* is undertaken to give a better understanding of a particular case for its own sake. In what is termed the *instrumental case study,* examination provides insight into an issue or a theory needing refinement. In this instance, the case takes a backseat, playing a supportive role and facilitating an understanding of the theory or issue.

During the *collective case study,* researchers move further away from any one particular case, studying a number of cases together as they inquire into the phenomenon or population at hand. In advance of the case study, researchers do not know whether the individual cases will manifest common characteristics. Their selection is based on the premise that understanding each individual case will increase knowledge about a larger group of cases. Whether case study researchers seek out what is particular about a case or what is common across cases, "the result is likely to be unique" (Stauffer, 1941, cited in Stake, 1994, p. 238).

Stake (1994) has argued that uniqueness of cases is likely to be pervasive, encompassing

- The nature of the case
- Its historical background
- The physical setting
- Other contextual factors, including economic, political, and legal realities
- Other cases through which this case is recognized
- Those informants through whom the case can be known

Issues of Generalization

Although Stake (1994, 1995, 2005) does not entirely disagree with Campbell (1975) that a case study may be a small step toward generalization, his opinion is that the commitment to generalizing (or building theory) may be damaging if it draws attention away from an understanding of the case itself. Indeed, some case study researchers deliberately do not seek causal determinations, mainly because their epistemology features coexisting events and conditions more than explanatory premises.

Source of Values in Case Studies

Perhaps Stake's most unusual claim (1994, 1995) is that not only the values of stakeholders but also the interpretations to be made by readers should be deliberately honored. The evaluator, according to Stake (1994, 1995), is a unique observer, interacting with a unique evaluand to assist unique readers in understanding that entity in its context. In spite of the potential for multiple, differing values and views, including those of the evaluator, Stake (1994, 1995) has expressed confidence that a wide range of readers would find such diversity of views useful.

Stake's View of Validating Case Study Evaluations

Stake (1994) believes it must be the researcher who tells his or her version of the case's story while retaining sound empathy toward the object of observation and honoring the views of its diverse group of stakeholders. As he said, "More will be pursued than was volunteered. Less will be reported than was learned" (Stake, 1994, p. 240). Case study research, according to Stake (1994), should provide grounds for validating reported statements, including a clear account of any generalizations that were made. Use of the concept of triangulation will prove essential to case study researchers, for their goal is understanding the case from a range of relevant perspectives, as well as seeking to minimize misinterpretations.

Stake's View of Case Study Methods

The methods of study supported by Stake (1994) are the use of the most intelligent observers and observation techniques possible and, underlying this, reflection: "Qualitative case study is characterized by the main researcher spending substantial time on-site, personally in contact with activities and operations of the case, reflecting, revising meanings of what is going on" (Stake, 1994, p. 242). Along another line, Stake (1994, 1995, 2005) has urged qualitative researchers to be aware of ethical considerations to protect human subjects.

It is not possible in this short account to give more than a brief outline of the case study research methodology that Stake (1995) put forward. Differing from the mainstream of case study evaluators, Stake (1994, 1995) has emphasized that the impression, held by some, of a case study as simply sharp observation is not very useful. A range of disciplines is needed, including designing good questions, organizing concepts, developing a cognitive framework to guide data gathering, and planning structures for appropriate presentation of interpretations to others. (For a clear and cogent description of these and other methodological elements of Stake's approach [1995] to case study research, we refer readers to Chapters 2, 4, 5, and 7 in *The Art of Case Study Research*.)

Stake's Use of Experience to Develop Evaluation Theory

Much of Stake's theory development over the years has been based on field experience. By contrast, some theory developers of the naturalistic or relativistic inclination have given little, if any, indication in their writing that they have experienced more than a modicum of fieldwork.

Moreover, as Stake (1995) has stressed, the execution of a particular case study affords a unique opportunity to learn, and learning from on-site situations is essential for the development of credible theory.

Stake's View of Qualifications to Conduct Case Study Evaluations

Stake (1994) summed up what he considered the major conceptual responsibilities of the qualitative case study researcher (p. 244):

1. Bounding the case and conceptualizing the object of study
2. Selecting phenomena, themes, or issues—that is, the research questions—to emphasize
3. Seeking patterns of data to develop the issues
4. Triangulating key observations and bases for interpretation
5. Selecting alternative interpretations to pursue
6. Developing assertions or generalizations about the case

Through his work with case studies, Stake (1994, 1995, 2005) has concluded that applications of this approach can help in refining theory, and, by the very nature of the complexities the approach engenders, can suggest limitations to generalizability of reported findings. Fundamentally, understanding of the individual case, not generalization, remains the purpose of the case study. The approach is embedded in personal discipline, and its success is determined by this factor.

Case Study Research: The Views of Robert Yin

Although most of this chapter is dedicated to Stake's case study approach (1994, 1995, 2005), it is useful to briefly contrast his views on case studies with those put forth by Yin (1992, 1998, 2009), another pioneer of applying case study methodology to evaluations of programs. Unlike Stake, Yin (1998, 2009) has placed greater emphasis on generating causal knowledge through case study techniques, using a process of analytic generalization rather than statistical hypothesis testing (as is typical in many of the social sciences). In part, this emphasis on causal knowledge is manifest in Yin's rationale (2009) for selecting the case study as a method of inquiry:

> There's no formula, but your choice depends in large part on your research question(s). The more that your questions seek to explain some circumstance (e.g., "how" or "why" some social phenomenon works), the more that the case study method will be relevant. (p. 4).

In Yin's view (2009), a case study is "an empirical inquiry that investigates a contemporary phenomenon in depth and within its real-life context, especially when the boundaries between

phenomenon and context are not clearly evident" (p. 18). Moreover, according to Yin, a case study researcher often

> copes with the technical situation in which there will be many more variables of interest than data points, and as one result relies on multiple sources of evidence, with data needing to converge in a triangulating fashion, and as another result benefits from the prior development of theoretical propositions to guide data collection and analysis. (p. 18)

Yin's Preordinate and Multimethod Orientation

Unlike Stake (1994, 1995, 2005), who primarily views case study research as an emergent process, Yin (2009) has asserted that case study research should follow "a rigorous methodological path" (p. 3). Consequently, Yin (2009) views case study research and evaluation largely as a preordinate activity requiring careful planning, execution, data collection, and analysis (though he also recognizes that modification is sometimes necessary as a result of discovery during data collection, for example). In this regard, Yin (2009) has proposed numerous strategies to increase the rigor of case studies, including using technical case study protocols so as to increase reliability; using multiple sources of evidence and member checks to establish construct validity; pattern matching and explanation building to establish internal validity; and using theory in single-case studies and replication logic in multiple-case studies to establish external validity. In addition, Yin (2009) has stressed the importance of using mixed-method approaches within the context of case study research and evaluation rather than relying on a single method, whether qualitative or quantitative.

Yin's View of Different Types of Case Studies

Yin (2009), like Stake (1994), identified multiple types of case studies, each with differing purposes. Unlike Stake, however, Yin defined four basic case study designs, as shown in Figure 12.1. In each, the case or cases are embedded in a unique context. According to Yin (2009), case study designs can be *single case* or *multiple case* (the horizontal axis in the figure) as well as *holistic* or *embedded* in terms of their unit of analysis (the vertical axis in the figure). The resulting types of designs for case studies are *Type 1 single-case holistic* designs, *Type 2 single-case embedded* designs, *Type 3 multiple-case holistic* designs, and *Type 4 multiple-case embedded* designs.

As Yin (2009) noted, the primary distinction is between single- and multiple-case designs. The single-case design is appropriate when it represents a critical case, an extreme or unique case, a representative or typical case, a revelatory case (a study of a previously inaccessible phenomenon), or a longitudinal case (a study of the same case over two or more points in time), each of which addresses different types of questions. Single-case designs may also be developed according to whether the case study involves more than one unit of analysis. In a holistic design, there are no logical subunits within the case. Conversely, in an embedded design, other subunits within a case are of relevance and should be considered in the design

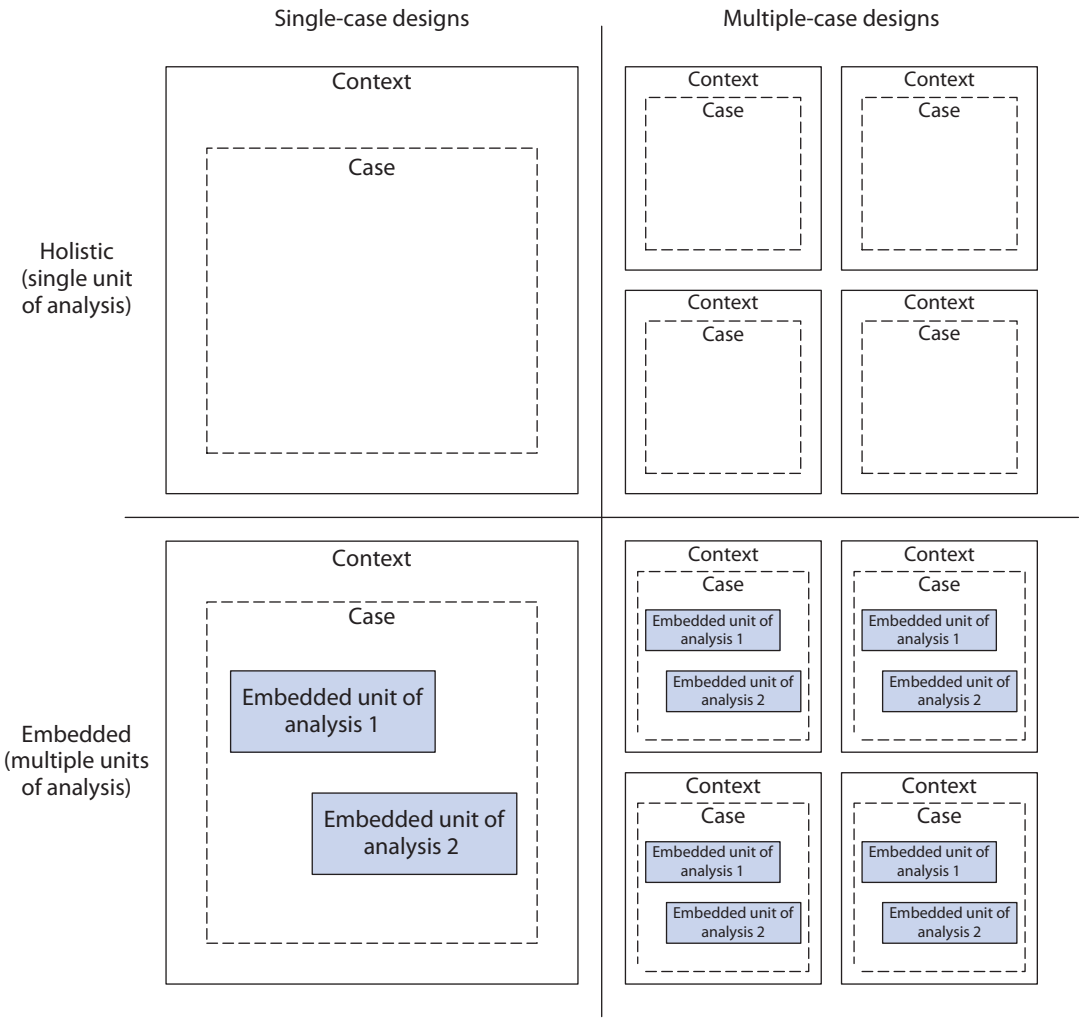

Figure 12.1 Basic Designs for Case Studies

Source: Yin, R. K. (2009). *Case study research: Design and methods* (4th ed.). Applied Social Research Methods Series, Vol. 5. Thousand Oaks, CA: Sage, 46.

of a case study (for example, a public program might consist of a large number of projects embedded within it).

The same general reasoning applies to multiple-case studies, except that the same study consists of more than one case. A typical example of multiple–case study research occurs when schools introduce innovations, such as new curricula, rearranged schedules, or new technologies, and some schools adopt only some of those innovations. In this scenario, each school might be the subject of an individual case study, but the study as a whole covers multiple cases (schools). As Yin (2009), noted, however, single- and multiple-case designs are "variants within the same methodological framework" (p. 53).

Yin's View of Needed Information

Yin (1998) encourages case study researchers and evaluators to gather one or more (preferably more) of six sources of evidence: documentation, archival records, interviews, direct observations, participant observations, and physical artifacts. Documentation can consist of letters, memorandums, e-mail correspondence, news clippings, proposals, progress reports, and other internal records, for example. Archival records include publically available data sources (such as U.S. census data), service records, budget and personnel records, maps and charts, and existing survey data. Interviews are considered one of the primary and most important data sources; they generally are semistructured rather than structured and should be fluid rather than rigid (Yin, 2009). Because case studies take place in the natural setting of the case, behaviors and environmental conditions often can be directly observed, either formally (for example, through structured observation protocols) or informally (for example, through casual observation). Participant observation, however, requires direct participation, rather than passive observation, in program activities, and evaluators may assume a variety of roles (such as program participant, program service provider, or program decision maker). Physical artifacts can include technological devices, tools or instruments, works of art, or other forms of physical evidence.

Yin's Advice on Analyzing and Interpreting Case Study Data

Yin (1992, 1998, 2009), perhaps more than any other writer on case study methodology, has made significant contributions to and advances in the analysis and interpretation of case study data. In particular, he has proposed five interrelated yet discrete approaches to the analysis of case study data: pattern matching, explanation building, time-series analysis, logic models, and cross-case synthesis.

Pattern matching, in which an empirically based pattern of results is compared to a predicted pattern of results, can be used to strengthen internal validity. Patterns can include nonequivalent dependent variables (see also Coryn & Hobson, 2011); rival explanations; and simpler patterns. Explanation building is a special type of pattern matching with the goal of building an explanation pertaining to the case.

Time-series analysis here is analogous to that used in experiments and quasi-experiments, involving the examination of patterns, over time, to build conclusions about a case or cases based on one or more dependent variables.

Logic models are used in case studies to stipulate a complex chain of events; observed events are then empirically matched to theoretically predicted events specified in a given logic model. Finally, Yin (2009) advises the use of cross-case synthesis when a case study consists of at least two cases. Here, the evaluator seeks out cross-case patterns through argumentative interpretation (that is, through logical reasoning based on empirical evidence) rather than by looking at numerical properties that cases have in common as supported by the case study data.

Particular Case Study Information Collection Methods

We have pointed out, as have Stake (1994, 1995, 2005) and Yin (1992, 1998, 2009), that using case study methodology, an evaluator can gather information about a particular phenomenon by a wide range of methods, both quantitative and qualitative. Whatever methods are used, the focus of the case study is the case itself. Practitioners have traditionally given particular emphasis to gathering qualitative information, particularly because both the quality and quantity of such information are likely to have fewer restrictions than that gathered quantitatively. However, much depends on the methods employed in the investigation, the case itself, the imagination and resourcefulness of the evaluator, and the kind of end information the client is seeking. Moreover, qualitative approaches that are carried out well should elicit information about a program's intended and unintended effects. Whatever methods are used, the aim of a case study is always to give as complete a picture as possible of the object being studied so that stakeholders may develop or enrich their understanding of the program and perhaps grasp the case study report's significance for decision making. The stronger the evaluator's skills of observation and reflection, the greater the understanding of the program among those involved in its progress and affected by its desired outcomes.

During the course of a case study, it is possible that planned methods may change form or new ones may be introduced according to the nature of the circumstances as they are illuminated. The evaluator must therefore be flexible and responsive to new or unusual circumstances, and must adapt to these as necessary. Record keeping is essential in case studies, even if the information will not be used in the final report. The necessity may not arise for a final report in the traditional sense; results may be conveyed to clients and others in many ways, with the emphasis always being on offering a clear and useful depiction of the program as revealed by the research. Because the report is based dominantly on qualitative information about a naturalistic setting, decisions about how findings are reported may evolve as the study progresses, with both intermediate and final reporting always being possibilities. Following is a brief description of some of the more commonly used qualitative methods in case study evaluations.

Documentation

Seeking to understand a program at multiple levels, as well as the holistic nature of a program, the evaluator logically should begin with an examination of existing documents, records, and other appropriate materials that give information about the program and characterize its geographical and organizational environment. Such records will give information about the program's personnel, processes, and progress. The evaluator should take notes about the key elements of each. Documents should give the perspectives of program stakeholders at various levels, and a lack of such important information must be noted for further investigation. Whatever the nature and specificity of records and other documents, the question should arise in the evaluator's mind about the kinds of research methods (most likely qualitative)

that should be employed in exploring these records. The perusal of documents and records has one other advantage: it should clearly delineate aspects of the program and thus save considerable and valuable evaluator time, which can more profitably be spent on searching out the information that is harder to obtain.

Content Analysis

In assessing documents and records, evaluators could find content analysis procedures to be valuable. Materials are analyzed and described as closely as possible, and processes and trends are noted. Content analysis as a data analysis method sharpens focus on significant aspects of programs. These are often exposed on the basis of their repetition within documents or any other useful emphases relevant to the program. Some of this information can be obtained qualitatively or quantitatively, depending on the kind of program knowledge that is presented and what is required. The important point is that the analyzer has in mind the questions to be answered. Quantitative content analysis depends on the development of coding units (such as words, paragraphs, or events), and these are then placed in categories. Although either stakeholders or the evaluator may select coding units and categories, however, the intent of the evaluation in the context of the wider spectrum of the object of the case study must be kept in mind.

Visits to the Program's Naturalistic Setting

We have emphasized the importance of an evaluator's experience and training when it comes to making sound professional judgments. This is particularly so for site visits involved in case studies. The main thrust of a case study evaluation is toward producing a qualitative, open-minded, in-depth inspection of a program. Although an evaluator may appropriately provide a program's stakeholders with formative feedback during a site visit, an equally important purpose of the site visit is to generate a rich, illuminating description of the program in its context and to render defensible summative conclusions. Such descriptions and conclusions should be grounded in a range of perceptions about and substantial evidence concerning the program within its environment. The insightful evaluator will use a whole range of methods during site visits, as part of careful advance planning, keen observations, and astute recording for later use. Planning will involve identifying the kinds of information that will be needed (while allowing that other, perhaps unexpected information may arise); preparing instruments, such as checklists, to be used on-site; working out the logistics of visits, including timing and personnel involved; and deciding on the kinds of reports that will be required in collaboration with the client. If a team is to undertake site visits, planning must extend to group meetings to discuss the allocation of duties and responsibilities and to ensure that the team is working toward a common end.

Observations

Central to the successful completion of any case study is the strength of observations. Often occurring during site visits, observations may include a discerning appraisal of interactions among personnel involved in a program, how and by what means the program is being

undertaken and developed (or how and why it is failing to develop), the strength of program leadership and delegation or otherwise of decision making, and the extent to which key stakeholders (those most affected by the program) are influenced by its evolving outcomes. Methods of observation for collecting relevant data and information may be quantitative or qualitative, although the latter are more likely to be appropriate, particularly when the observations are unstructured. Jorgensen (1989) pointed out the usefulness of unstructured observations during the early stages of a study as well as the importance of the evaluator's skills in selecting and delineating critical features of the case:

> 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 soon as possible and with the greatest possible detail because never again will you experience the setting so utterly unfamiliar. (p. 82)

As the study progresses, unstructured observations will continue in ways partly dictated by the kinds of information that unfold. If the use of observation is based on meetings with stakeholders (as it often is), an evaluator can obtain an increasingly useful depiction of the program, with all of its nuances, including the forces that prevail in the program's environment. If an evaluator is carrying out unstructured observations, an ongoing comparison of observations strengthens the understanding of the program's holistic nature.

Observations that are more structured are also essential and worth the time. Unlike unstructured observations that involve viewing aspects of a program in a general sense, structured observations focus on the program's idiosyncrasies; events associated with the program; a range of physical aspects; and particularly the interactions between leaders and other stakeholders, which have a strong influence on any program and should be carefully noted. Sensitive information may arise from observations. It is therefore most important that observations, particularly pertaining to personnel, be maintained securely and confidentially. The propriety standards developed by the Joint Committee on Standards for Educational Evaluation (1994) stress the importance of ethical behavior by evaluators in reaching and upholding agreements in such areas as anonymity and confidentiality.

Structured observations must be based on careful planning that includes attention to such items as observation scheduling; the kinds of instruments that will be used; and an appropriate time schedule for conducting observations, to be worked out in collaboration with program administrators. If a team is involved, participant training to ensure reliability among observations will be necessary.

Qualitative methods of observation usually focus on the observer's (or observers') viewing the interactions between group members objectively, while collecting information according to a prearranged schedule or checklist. The observer also can play a greater participatory role in group discussions, depending on the prevailing circumstances and the kinds of information sought. In such instances, it is usual for the evaluator to ask questions to help elucidate matters that have arisen during the observation period. Again, astute note taking and synthesizing of information will help build a more complete picture of the program with all its intricacies.

Interviewing

This area requires a high level of skill. Preparation for interviews is vital if they are to elicit the kinds of information that are sought to illuminate the program. By comparison with the use of questionnaires, conducting interviews can be a costly exercise, but one that is commonly employed to unravel some of a program's complexities and particularly stakeholders' reactions to these. Stakeholders' concerns about a program and their knowledge of it are perspectives that are essential for accurate and meaningful reporting in whatever form it might take. Much has been written, and will still be written, on successful methods for carrying out interviews. In all of this advice, the essential components are the experience of the interviewer, the degree of preparation, the importance of clearly understanding the program itself and the purpose for the interviewing, and the need to make the respondent feel at ease and useful to the study being undertaken. Once rapport between the interviewer and respondent has developed, the primary task of the interviewer is to listen and encourage discussion at a professional level.

Focus Groups

Focus groups are an extension of interviewing, involving groups of individuals who are closely connected with the subject program. Focus groups involve interactions between the interviewer and the group, and between group members themselves. Group members may be engaged to give their views on the case being studied or to react to a draft or final case study report. In the former situation, they may generate a great deal of useful information about the program, particularly if they have opposing or conflicting views on aspects of that program. In the latter situation, by reacting to a report they may help in such ways as assessing the case study's validity or identifying what they see as its implications for action. The interviewer's task is to make sure that dialogue remains focused on the topic under discussion. The more accurately participants relate their reactions to the program and other relevant experiences, the sharper the focus will be on desirable program changes that may be required. A number of factors come into play with this method. It is crucially important to select a set of participants representing a subset of stakeholders that is appropriate to the focus group's charge. Idiosyncratic beliefs and value systems are never far from the surface during focus groups. Any attitudes the participants hold may influence the program's progress and, in many instances, its success or failure. Focus groups, properly constituted and conducted, certainly add very useful dimensions to a case study evaluation.

Summary

The case study approach is appropriate in program evaluation, particularly because it requires no control of treatments, subjects, or programs in their naturalistic setting. In addressing focal issues, evaluators using such an approach triangulate multiple perspectives using a range of methods (according to the needs of each unique situation) and sources of information. From close contextual investigations of influences on the subject program, the case study evaluator progresses to a holistic, in-depth assessment of the program, with its complexities

and human interplay. Although it is possible to undertake case studies retrospectively on the basis of recorded data and documents, case studies are more likely to occur in real time. Case study methodology has become increasingly useful to evaluators as investigators and to administrators and other stakeholders seeking an accurate depiction of a program. This chapter has provided our perspective on the case study approach along with the perspectives of two authors—Stake and Yin—who have contributed substantially to the application of case study methods to program evaluations.

REVIEW QUESTIONS

1. What are some of the significant differences between experimental design and case study methods?
2. There are marked differences between Brinkerhoff's Success Case Method (see Chapter 6) and the case study approach. State these differences.
3. What are the primary differences between Stake's and Yin's approaches to case studies? What are the similarities?
4. What do you understand by the terms *structured observation* and *unstructured observation*?
5. How would you use content analysis to delineate salient issues in an examination of documents and records?
6. What qualifications should evaluators possess, and what main procedures should they follow, to carry out successful case study evaluations?
7. You have been commissioned to carry out a case study of a fourth-grade music appreciation program. Following Yin's advice, which of the following two questions is the more appropriate?
   a. What can I learn from this single case?
   b. What qualitative methods would be most effective for generating unequivocal, generalizable information about this program as a representative of similar programs?

   Give reasons for your choice of question and why you rejected the alternative. Also, discuss whether Stake would be likely to agree or disagree with what you see as Yin's response to this question.
8. Discuss whether a quantitative approach to a case study would be more appropriate in evaluating a weight-loss program than in evaluating the music appreciation program referenced in question 7.
9. What does Stake mean by intrinsic, instrumental, and collective case studies?
10. Stake maintained that there are six ways in which a program is unique, and that researchers are required to gather data on all of these. Identify and briefly define these six unique characteristics.

Group Exercises

This chapter has provided an abbreviated overview of two approaches to case study methods for evaluation. We hope that your discussions will enhance your understanding of some of the differences (and similarities) between the two approaches to case studies.

Exercise 1

According to Stake, a case study is defined not by a methodology, but by the choice of object to be studied. Discuss this assertion along these lines:

- Stake's view on what should be the basic intent of a case study
- The general types of evaluation methods (if any) he views as appropriate to a particular study
- How a choice of object is to be defined

Exercise 2

Is Stake unrealistic in stating that a case study does not need to be bound by time? In your response, refer to evaluation situations with which you are familiar.

Exercise 3

Summarize your understanding of Yin's position in regard to the appropriateness of a preordinate and multimethod orientation to case study evaluations. Then consider how this orientation does or does not align with Stake's recommended approach to case study evaluations. Finally, discuss whether Stake and Yin agree or disagree concerning the need for triangulation of findings.

Suggested Supplemental Readings

Campbell, D. T. (1975). Degrees of freedom and the case study. *Comparative Political Studies*, *8*, 178–193.

Jorgensen, D. L. (1989). *Participant observation: A methodology for human studies*. Thousand Oaks, CA: Sage.

Stake, R. E. (1988). Seeking sweet water. In R. Jaeger (Ed.), *Complementary methods for research in education* (pp. 253–300). Washington, DC: American Educational Research Association.

Stake, R. E. (1994). Case studies. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 236–247). Thousand Oaks, CA: Sage.

Stake, R. E. (1995). *The art of case study research*. Thousand Oaks, CA: Sage.

Stake, R. E. (2005). *Multiple case study analysis*. New York, NY: Guilford Press.

Stauffer, S. (1941). Notes on the case study and the unique case. *Sociometry*, *4*, 349–357.

Yin, R. K. (1992). The case study as a tool for doing evaluation. *Current Sociology*, *40*(1), 121–137.

Yin, R. K. (1998). The abridged version of case study research: Design and method. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 229–260). Thousand Oaks, CA: Sage.

Yin, R. K. (2009). *Case study research: Design and methods* (4th ed.). Applied Social Research Methods Series, Vol. 5. Thousand Oaks, CA: Sage.

CHAPTER 13

DANIEL STUFFLEBEAM'S CIPP MODEL FOR EVALUATION

An Improvement- and Accountability-Oriented Approach

Overview of the Chapter

The CIPP evaluation model is a comprehensive framework for conducting formative and summative evaluations of programs, projects, personnel, products, organizations, policies, and evaluation systems. Basically, the model provides direction for assessing context (in terms of an enterprise's need for corrections or improvements); inputs (strategies, operational plan, resources, and agreements for proceeding with a needed intervention); process (the intervention's implementation and costs); and products (the effort's positive and negative outcomes).

This chapter summarizes the CIPP model's roots; lists the model's various applications across a wide range of sectors in society; defines context, input, process, and product evaluations; defines evaluation in general and other key concepts associated with the model; analyzes the model's formative and summative uses; presents the model's philosophical stance and code of ethics; emphasizes the model's focus on improvement; discusses its values component; delineates relevant procedures; and explains and illustrates the model's systems orientation.

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **The historical roots and range of applications of Daniel Stufflebeam's context, input, process, and product (CIPP) model**
- **The CIPP model's conceptual and operational definitions of evaluation**
- **The CIPP model's grounding in professional standards for evaluations**
- **The CIPP model's conceptual and operational definitions of context, input, process, and product evaluation**
- **The CIPP model's formative and summative uses**
- **The CIPP model's approach to engaging and serving stakeholders**
- **The CIPP model's values, improvement, objectivist, and systems orientations**
- **The CIPP model's requirement for and approach to obtaining metaevaluations**

CIPP Model in Context

Understanding the CIPP model's background is important for confirming the model's significance based on its original development to address national program evaluation needs, its widespread use ever since, its evolution based

on lessons from applications, and its integrity owing to adherence to what the evaluation field has defined as sound evaluation practice.

Roots of the CIPP Model

The CIPP model was created in the late 1960s to help improve and achieve accountability for federally funded U.S. public school projects, especially those keyed to improving teaching and learning in inner-city school districts. Over the years, the model has been further developed. It has been adapted and applied in the United States and many other countries and across a wide range of disciplines and service areas.

The CIPP model is based on learning by doing—that is, an ongoing effort to identify and correct mistakes made in evaluation practice, to invent and test needed new procedures, and to retain and incorporate especially effective practices. The history of the model's development parallels and is a main part of the history of development of evaluation models and procedures since the mid-1960s. Like other new approaches to evaluation, it was created because the classic evaluation approaches of experimental design, objectives-based evaluation, peer or expert review site visits, and standardized achievement testing proved to be of limited use and often unworkable and even counterproductive for evaluating emergent federal programs in dynamic social contexts and particularly public school districts. (A detailed account of the model's development appears in the second edition of Alkin's *Evaluation Roots* [2013]).

The early work in developing the model is documented in *Educational Evaluation and Decision Making* (Stufflebeam et al., 1971), a book produced by a national study committee on evaluation that Phi Delta Kappa International (PDK) appointed in 1969. The book's authors sharply criticized the traditional views of evaluation, analyzed the evaluative information needs in decision making, elaborated the CIPP model, closely examined the problems of multilevel evaluation, addressed the issue of institutionalizing systematic evaluation, discussed the need for evaluation training, and proposed that criteria for judging evaluations should include utility and feasibility as well as technical adequacy.

An important lesson in regard to criteria for assessing evaluations was that evaluations can go very wrong if they are keyed exclusively to criteria of technical adequacy, such as the requirements for internal and external validity being promulgated in the 1960s for judging experiments (Campbell & Stanley, 1963; Shadish, Cook, & Campbell, 2002). The PDK book's disaggregation of utility criteria into relevance, importance, timeliness, clarity, and credibility, plus its recommendation that there be a prudential criterion concerned with conserving resources, was a precursor of work the Joint Committee on Standards for Educational Evaluation (1981, 1988, 1994, 2003, 2009, 2011) would do in defining standards for evaluations of utility, feasibility, propriety, accuracy, and evaluation accountability.

The Model's Range of Applications

The model is adaptable for application by a wide range of users, including evaluators, program specialists, researchers, developers, policy groups, leaders, administrators, committees or task groups, and laypersons. In the last case, one could use the CIPP model to guide and assess a home remodeling project. Context evaluation would be employed to assess a home's adequacy for

meeting the occupants' needs and to identify problems requiring attention (such as a need for landscaping, painting, replacement shingles, new gutters, new flooring, added insulation, new windows, rewiring, new plumbing lines, new appliances, an enlarged kitchen, an exterior deck, termite protection, and so on), and then for deciding on which improvements to pursue. Input evaluation would be used to obtain and assess alternative architectural and contractor-produced plans, to determine the quality and costs of house rehabilitation products and services, and also to contract and budget for the selected products and services. Through process evaluation, the home owner (often with support from licensed inspectors) would monitor and take steps to ensure quality, safety, cost containment, on-time performance, and adequate cleanup and follow-up. Ultimately, the home owner would employ product evaluation to assess the results of the remodeling project, obtain needed inspections to ensure compliance with relevant codes and laws and with the original contractual agreements, and then pay the bills.

As this example illustrates, the CIPP model is a commonsense approach to ensuring cost-effectiveness in starting, planning, carrying out, and completing needed improvement efforts. It applies to one's day-to-day decisions and actions as well as to complex enterprises operated by private and public organizations. Beyond guiding the development and implementation of improvement efforts, the model is also configured to meet an enterprise's post hoc accountability requirements.

Guili Zhang of East Carolina University indicated in personal correspondence that her search for relevant literature on the CIPP model identified about 200 CIPP-related evaluation studies, journal articles, and doctoral dissertations spanning many nations and fields. She found that the model had been applied in 134 doctoral dissertations at eighty-one universities involving a total of thirty-nine disciplines. She also cited a sample of 55 published studies (among many more such studies) that employed the model in such disciplines as agriculture, aviation; business; communication; distance education; elementary, tertiary, and secondary education; government; health care; international development; law; philanthropy; psychology; religion; and sociology. Those employing or contracting others to employ the model included government officials, foundation officers, program and project staff, international assistance personnel, agricultural extension agents, school administrators, church officials, physicians, nurses, military leaders, and evaluators.

In this book's first edition (Stufflebeam & Shinkfield, 2007), the CIPP model chapter cited some areas of application beyond those identified by Zhang: productivity of private colleges and historically black colleges; community programming for youth; community and economic development; house construction and rehabilitation; and systems for evaluating teachers, administrators, and military personnel. Clearly, use of the CIPP model is widespread.

Pertinent references to development and application of the model include Adams (1971); Candoli, Cullen, and Stufflebeam (1997); Gally (1984); Guba and Stufflebeam (1968); Nevo (1974); Stufflebeam (1966b, 1967, 1969, 1971b, 1983, 1985, 2003a, 2003b, 2004b, 2005, 2007); Stufflebeam et al. (1971); Stufflebeam and Webster (1988); Webster (1975); and Zhang et al. (2010). A detailed checklist for applying the CIPP evaluation model is available at www.wmich.edu/evalctr/checklists, and an updated version is available from the Jossey-Bass Web site supporting this book at www.josseybass.com/go/evalmodels.

Conceptual and Operational Definitions of Evaluation

The CIPP model is grounded in general and operational definitions of evaluation, main uses of evaluations, and professional standards for guiding and judging evaluations. Generally, an evaluation is a systematic investigation of some object's value. Operationally, evaluation is the process of delineating, obtaining, reporting, and applying descriptive and judgmental information about an object's value, as defined by such criteria as quality, worth, probity, equity, feasibility, cost, efficiency, safety, and significance.

Professional Standards

Professional standards for evaluations are principles commonly agreed to by specialists in the conduct and use of evaluations for determining an evaluation's utility, feasibility, propriety, accuracy, and evaluation accountability. Basically, the CIPP model is an organized approach to meeting the evaluation profession's standards as defined by the Joint Committee (1981, 1988, 1994, 2003, 2009, 2011). This book's Chapter 3 on evaluation standards is an important supplement to this chapter.

Overview of the CIPP Categories

The CIPP model's core concepts are evaluations of an entity's context, inputs, processes, and products, as denoted by the letters of the acronym.

In context evaluations, evaluators assess needs, problems, assets, and opportunities, plus relevant contextual conditions and dynamics. Decision makers use context evaluations to define goals and set priorities and to make sure program goals are targeted to address significant, assessed needs and problems. Oversight bodies and program stakeholders use context evaluation findings to judge whether the program was guided by appropriate goals and also to judge outcomes for their responsiveness to the program's targeted needs, problems, and goals.

In input evaluations, evaluators assist with program planning by identifying and assessing alternative approaches and subsequently assessing procedural plans, staffing provisions, and budgets for their feasibility and potential cost-effectiveness in regard to meeting targeted needs and achieving goals. Decision makers use input evaluations to identify and choose among competing plans, write funding proposals, allocate resources, assign staff, schedule work, and ultimately help others judge an effort's plans and budget.

In process evaluations, evaluators monitor, document, assess, and report on the implementation of program plans. Such evaluators provide feedback throughout a program's implementation and later report on the extent to which the program was carried out as intended and required. Program staff use periodic process evaluation reports to take stock of their progress, identify implementation issues, and adjust their plans and performance to ensure program quality and on-time delivery of services. At the end of the program or after a program cycle, the program's staff, overseers, and constituents may use the process evaluation's documentation to judge how well the program was carried out. They may also use this documentation to judge whether a program's possibly deficient outcomes were due

to a weak intervention strategy or to inadequate implementation of the strategy. In addition, the program approach's potential adopters may seek out and use the findings of the process evaluation to guide their adaptation and application of the approach.

In product evaluations, evaluators identify and assess costs and outcomes—intended and unintended, short term and long term. They provide feedback during a program's implementation on the extent to which program goals are being addressed and achieved. At the program's end, product evaluation helps identify and assess the program's full range of accomplishments. Program staff use interim product evaluation feedback to maintain focus on achieving important outcomes and to identify and address deficiencies in the program's progress toward achieving important outcomes. Ultimately, product evaluations involve assessing and reporting on a program's unintended as well as intended outcomes. Program overseers, funders, and constituents use final product evaluation results to judge whether the program's accomplishments were significant and worth the cost. The program's potential adopters would use product evaluation findings as the most important information for deciding whether to adopt the program. Product evaluation's key questions are: Did the program achieve its goals? Did it successfully address the targeted needs and problems? What were the program's side effects? Were there negative as well as positive outcomes? Were the program's accomplishments worth the cost?

In summing up long-term evaluations, the product evaluation component may be divided into four subparts of assessment: reach to the targeted beneficiaries, effectiveness, sustainability, and transportability. These product evaluation subparts necessitate asking, Were the right beneficiaries reached? Were the targeted needs and problems addressed effectively? Were the program's accomplishments and the mechanisms to produce them sustained and affordable over the long term? Did the strategies and procedures that produced the accomplishments prove to be transportable, adaptable, and affordable for effective use elsewhere?

Formative and Summative Uses of Context, Input, Process, and Product Evaluations

Main uses of evaluations, based on the CIPP model, are to guide and strengthen enterprises; issue accountability reports; help disseminate effective practices; increase understanding of the involved phenomena; and, as appropriate, make decision makers, stakeholders, and consumers aware of evaluands that proved unworthy of further use.

Consistent with its improvement focus, the CIPP model places priority on guiding planning and implementation of development efforts. In the model's formative role, context, input, process, and product evaluations, respectively, ask: What needs to be done? How should it be done? Is it being done? Is it succeeding? Prior to and during the decision-making and implementation processes, the evaluator submits reports addressing these questions to help guide and strengthen decision making, keep stakeholders informed about findings, help staff work toward achieving successful outcomes, and help them maintain an accountability record. In this vein, the model's intent is to supply evaluation users—such as policy boards, administrators, and program staffs—with timely, valid information of use in identifying an

appropriate area for development; formulating sound goals, activity plans, and budgets (often associated with development or improvement efforts); successfully carrying out work plans; strengthening existing programs or services; periodically deciding whether and, if so, how to repeat or expand an effort; disseminating effective practices; contributing to knowledge in the area of service; and meeting a financial sponsor's accountability requirements.

The model also includes a requirement and provides direction for conducting retrospective, summative evaluations to serve a broad range of stakeholders. Possible stakeholders include funding organizations, persons receiving or considering using the sponsored services, policy groups and program specialists outside the program being evaluated, and researchers. In preparing the summative report, the evaluator refers to the store of formative context, input, process, and product information and obtains other needed information. The evaluator uses this information to address the following retrospective questions: Was the program (or other evaluand) keyed to clear goals based on the assessed needs of beneficiaries? Was the effort guided by a defensible procedural design; a functional staffing plan; an effective and appropriate process of stakeholder involvement; and a sufficient, appropriate budget? Were the plans executed competently and efficiently and modified as needed? Did the effort succeed, in what ways and to what extent, and why or why not? Potential consumers need answers to such summative questions to help assess the quality, cost, utility, and competitiveness of programs, products, or services they might adopt or acquire and use. Other stakeholders might want evidence on the extent to which their tax dollars or other types of support yielded responsible actions and worthwhile outcomes. If evaluators effectively conduct, document, and report on formative evaluations, they will have much of the information needed to produce a defensible summative evaluation report. Such information will prove invaluable to both internal and external evaluators with an assignment to summatively evaluate a project, program, service, or other entity.

Table 13.1 summarizes uses of the CIPP model for both formative and summative evaluations. The matrix's eight cells (formative and summative roles of context, input, process, and product evaluation) encompass much of the evaluative information required to guide enterprises and produce credible, and therefore defensible, summative evaluation reports.

Philosophy and Code of Ethics Underlying the CIPP Model

The CIPP model has a strong orientation toward service and the principles of a free society. It calls for evaluators and clients to identify and involve rightful beneficiaries; clarify the forms of assistance they need; obtain information of use in designing responsive programs and other areas of assistance; assess and help guide the intervention's effective implementation; and ultimately assess the intervention's value (for example, its quality, worth, probity, equity, feasibility, cost, efficiency, safety, and/or significance). The thrust of CIPP evaluations is to provide sound information and judgments that will help service providers regularly assess and improve services and make effective and efficient use of resources, time, and technology to appropriately and equitably serve the well-being and targeted needs of rightful beneficiaries.

Table 13.1 Relevance of Four Evaluation Types to Formative and Summative Evaluation Roles

	Types of Evaluation			
Evaluation Role	**Context Evaluation**	**Input Evaluation**	**Process Evaluation**	**Product Evaluation**
Formative evaluation: Prospective application of CIPP information and judgments to assist with decision making, program implementation, quality assurance, and accountability	Providing guidance for identifying needed interventions, choosing goals, and setting priorities *by assessing and reporting on needs, problems, assets, and opportunities*	Providing guidance for choosing a program strategy and settling on a sound general implementation plan and budget *by assessing and reporting on alternative strategies and resource allocation plans and subsequently closely examining and judging the specific operational plan*	Providing guidance for implementing the operational plan *by monitoring, documenting, judging, and repeatedly reporting on program activities and expenditures*	Providing guidance for continuing, modifying, adopting, or terminating the program *by identifying, assessing, and reporting on intermediate and longer-term outcomes, including side effects*
Summative evaluation: Retrospective use of CIPP information to sum up the program's value (for example, its quality, worth, probity, equity, feasibility, cost, efficiency, safety, and/or significance)	Judging goals and priorities *by comparing them to assessed needs, problems, assets, and opportunities*	Judging the implementation plan and budget *by comparing them to the targeted needs of intended beneficiaries, contrasting them with those of critical competitors, and assessing their compatibility with the implementation environment*	Judging program implementation *by fully describing and assessing the actual processes and costs, plus comparing the planned and actual processes and costs*	Judging the program's success *by comparing its outcomes and side effects to targeted needs, examining its cost-effectiveness, and (as feasible) contrasting its costs and outcomes with those of competitive programs; also by interpreting results against the effort's outlay of resources and the extent to which the operational plan was both sound and effectively executed*

Involving and Serving Stakeholders

CIPP evaluations must be grounded in the democratic principles of equity and fairness. A key concept used in the model is that of stakeholders: those who are intended to use the findings, those who may otherwise be affected by the evaluation, and those expected to contribute to the evaluation. Consistent with the Joint Committee's *Program Evaluation Standards* (1994, 2011), the evaluator should search out all relevant stakeholder groups and engage at least their representatives in hermeneutic and consensus-building processes. He or she should engage them to help with affirming and clarifying foundational values, defining evaluation questions, clarifying evaluative criteria, obtaining needed information, interpreting findings, assessing evaluation reports, and disseminating and using findings.

Because information empowers those who hold it, the CIPP model emphasizes the importance of evenhandedness in involving and keeping informed all of a program's stakeholders. Moreover, evaluators should strive to reach and involve those most in need and with little access to and influence over services. Although evaluators should control the evaluation process to ensure its integrity, CIPP evaluations accord beneficiaries and other stakeholders more than a passive recipient's role. Evaluators are expected to keep stakeholders informed and provide them with appropriate opportunities to contribute. Involving all levels of stakeholders is considered ethically responsible because it equitably empowers the disadvantaged as well as the advantaged to help define the appropriate evaluation questions and criteria; provide evaluative input; critique draft reports; and receive, review, and use evaluation findings. Involving all stakeholder groups is also wise because sustained, consequential involvement positions stakeholders to contribute information and valuable insights and inclines them to study, understand, accept, value, and act on evaluation reports.

Improvement Orientation

A fundamental tenet of the CIPP model is that evaluation's purpose is not only to prove but also—and more important—to improve. Evaluation is thus conceived first and foremost as a functional activity oriented in the long run to stimulating, aiding, and abetting efforts to strengthen and improve enterprises. The model also posits that some programs or services will prove unworthy of attempts to improve them or overly expensive and should be discredited or terminated. By helping stop unneeded, unsustainable, corrupt, or hopelessly flawed efforts, evaluations can and should serve an improvement function by helping organizations to free up resources and time for worthy efforts.

In the first author's experience, based on evaluating government programs over several decades, some federal programs have been shown to be culpable for fraudulent, wasteful, or inappropriate use of federal funds. Unfortunately, too often such malfeasance has not been exposed, corrected, and (as appropriate) penalized. Evaluation has an important role, not only in uncovering fraud, waste, and abuse in assessed programs but also in advocating for corrective action by appropriate authorities.

Objectivist Orientation

The CIPP model's epistemological orientation is objectivist, not relativistic. Objectivist evaluations are based on the theory that moral good is objective and independent of personal or simply human feelings. Evaluators conducting such evaluations are doing work that is firmly grounded in ethical principles, such as those in the United Nations' Universal Declaration of Human Rights and the U.S. Bill of Rights; they strive to control bias, prejudice, and conflicts of interest in conducting assessments and reaching conclusions; they invoke and justify appropriate and (where they exist) established technical standards of quality; they obtain and validate findings from multiple sources; they search for best answers, although these may be difficult to find; they set forth and justify best available conclusions about the evaluand; they report findings honestly, fairly, and as circumspectly as necessary to all right-to-know audiences;

they subject the evaluation process and findings to independent assessments against pertinent standards; and they identify needs bearing further investigation. Fundamentally, objectivist evaluations are intended over time to lead to conclusions that are correct—not correct or incorrect relative to an evaluator's or other party's predilections, position, preferences, or point of view. The model contends that when different objectivist evaluations are focused on the same object in a given setting, when they are keyed to fundamental principles of a free society and to agreed-on criteria of merit, when they involve meaningful engagement of all stakeholder groups in the quest for answers, and when they conform to the evaluation field's standards, different, competent evaluators will arrive at fundamentally equivalent, defensible conclusions.

Standards and Metaevaluation

The model calls for evaluators to meet the professional standards of evaluation and subject their evaluations to both formative and summative metaevaluations. The main standards invoked in the model require evaluations to meet professionally defined requirements for utility, feasibility, propriety, accuracy, and evaluation accountability. At a minimum, evaluators should conduct their own formative and summative metaevaluations. An evaluator should use the formative metaevaluation to guide the evaluation work and correct deficiencies along the way. In the final evaluation report, the evaluator should state and explain his or her judgments of the extent to which the evaluation met each of the relevant Joint Committee (2003, 2009, 2011) standards. As feasible, the evaluation should be subjected to an external, independent metaevaluation. Preferably, a party other than the evaluator, such as the client or a private foundation, should choose and fund the external metaevaluator. This helps avoid any appearance or fact of the evaluator's possible conflict of interest's having influenced the content of the external metaevaluation report. The external metaevaluator's report should be made available to all members of the right-to-know audience. (Further information on these points appears in Chapter 3 (on professional standards) and Chapter 25 (on metaevaluation).

The Model's Values Component

Figure 13.1 summarizes the basic elements of the CIPP model in three concentric circles and portrays the central importance of defined values. The inner circle denotes the core values that should be defined and used to undergird a given evaluation. The wheel surrounding the values is divided into four evaluative foci associated with any program or other endeavor: goals, plans, actions, and outcomes. The outer wheel indicates the type of evaluation that serves each of the four evaluative foci: context, input, process, or product evaluation. Each two-directional arrow represents a reciprocal relationship between a particular evaluative focus and a type of evaluation.

The goal-setting task raises questions for a context evaluation, which in turn provides information for validating or improving goals. Planning improvement efforts generate questions for an input evaluation, which correspondingly provides judgments of plans and direction for strengthening plans. Program actions bring up questions for a process evaluation,

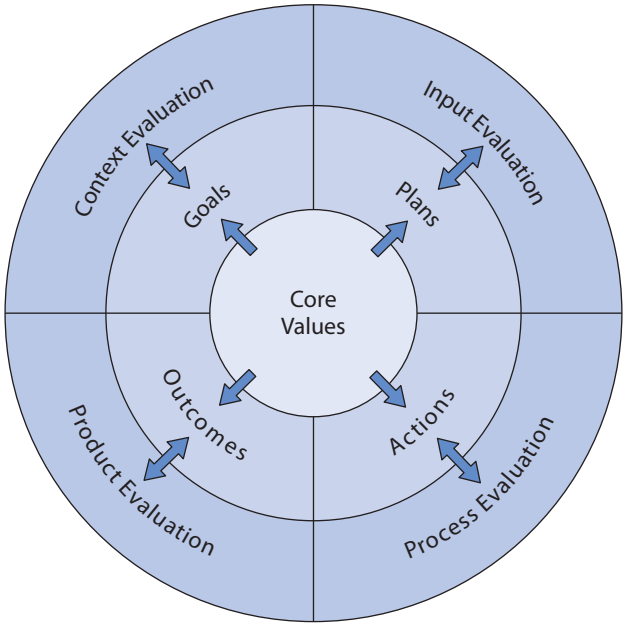

Figure 13.1 Key Components of the CIPP Evaluation Model and Associated Relationships with Programs

which provides judgments of activities plus feedback for strengthening staff performance. Accomplishments, lack of accomplishments, and side effects command the attention of a product evaluation, which ultimately yields judgments of outcomes and helps identify needs for achieving better results.

These relationships are made functional by grounding evaluations in core values, referenced in the scheme's inner circle. Evaluation's root term is *value*. This term refers to any of a range of ideals held by a society, group, or individual. The CIPP model calls for the evaluator and client to identify and clarify the values that will undergird a particular evaluation. Examples of educational values—applied in evaluations of U.S. public school programs—are success in helping all students meet a state's mandated academic standards; helping all children develop basic academic skills; helping each child fulfill her or his potential for educational development; aiding and reinforcing the development of students' special gifts and talents; upholding human rights; meeting the needs of children with disabilities and underprivileged children; developing students as good citizens; ensuring equality of opportunity; effectively engaging parents in the healthy development of their children; nurturing and developing the school's primary resource, its teachers; attaining excellence in all aspects of schooling; conserving and using resources efficiently; ensuring safety of products and procedures and of students and staff; maintaining separation of church and state; employing research and innovation to strengthen teaching and learning; and maintaining accountability. Essentially evaluators should take into account a set of pertinent societal, institutional, program, and professional and technical values when assessing programs or other entities.

The values provide the foundation for deriving or validating particular evaluative criteria. Selected criteria, along with stakeholders' questions, help clarify an evaluation's information

needs. These in turn provide the basis for selecting and constructing the evaluation instruments and procedures, accessing existing information, collecting new information, and defining interpretive standards.

Also, a values framework provides a well-knit point of reference for detecting unexpected defects and strengths. For example, through broad, values-oriented surveillance, an evaluator might discover that a program excels in meeting students' targeted academic needs but has serious deficiencies, such as racist practices, unsafe equipment, teacher burnout, waste of resources, or graft. On the positive side, examination of a program against a backdrop of appropriate values might uncover unexpected positive outcomes, such as strengthened community support of schools, invention of better teaching practices, improved teacher morale, or more engaged and supportive parents.

Using the CIPP Framework to Define Evaluation Questions

Table 13.2 illustrates how the CIPP framework can be used to identify and review possible questions for a program evaluation. The first row identifies generic questions for formative evaluations of context, inputs, processes, and products, and analogously the second row identifies generic questions for summative evaluations. This chart is offered for the evaluator's use in engaging the client and other stakeholders in the process of identifying and defining the questions that will guide a particular evaluation.

Delineation of the CIPP Categories and Relevant Procedures

This section provides a more specific discussion of each type of evaluation. Table 13.3 is a convenient overview of the essential meanings of context, input, process, and product evaluation. It defines these four types of studies according to their objectives, methods, and uses. This section also describes certain techniques that evaluators have found useful for conducting each type of evaluation. No one evaluation is likely to use all of the techniques referred to here. They are presented to give an idea of the range of qualitative and quantitative methods that are potentially applicable in CIPP evaluations.

Context Evaluation

An evaluator employs a context evaluation to assess needs, problems, assets, and opportunities within a defined environment. Needs include those things that are necessary or useful for fulfilling a defensible purpose. Problems are impediments to overcome in meeting and continuing to meet targeted needs. Assets include accessible expertise and services, usually in the local area, that could be used to help fulfill the targeted purpose. Opportunities especially include funding sources that might be tapped to support efforts to meet needs and solve associated problems. Defensible purposes define what is to be achieved related to the institution's mission while adhering to ethical and legal standards.

Although context evaluation is often referred to as needs assessment, the latter term is too narrow in that it focuses on needs and omits concerns about problems, assets, and opportunities. All four elements are critically important in designing sound programs, projects,

Table 13.2 Illustrative Evaluation Questions

Evaluation Role	Types of Evaluation			
	Context Evaluation	**Input Evaluation**	**Process Evaluation**	**Product Evaluation**
Formative	• What are the highest-priority needs in the program area of interest? • What goals should be pursued to meet the needs?	• What are the most promising approaches to meeting the targeted needs and goals? • How do these approaches compare in regard to potential success, costs, and so on? • How can the most promising approach be effectively designed, staffed, funded, and implemented? • What might be some barriers to effective implementation?	• To what extent is the funded program proceeding on time, within budget, and effectively? • If needed, how can the program's design be improved? • How can implementation be strengthened?	• To what extent are indicators of success being observed and assessed? • What other indicators, if any, have emerged that show that the program is or is not succeeding? • What side effects (positive or negative) are emerging? • How should implementation be modified to sustain success?
Summative	• To what extent did this program area address high-priority needs? • To what extent did program goals reflect the targeted assessed needs?	• What approach to goal achievement was chosen? • How did the chosen strategy compare to other viable approaches in regard to prospects for success, feasibility, and costs? • How well was the chosen strategy converted to a sound, feasible work plan?	• To what extent was the program carried out as planned or modified with an improved plan? • How well was the program executed? • What was the program's overall cost?	• To what extent did this program effectively address the original assessed needs and achieve its goals? • Were there any unanticipated negative or positive side effects? • What conclusions can be reached concerning the program's cost-effectiveness, sustainability, and broad applicability?

and services and should be considered in context evaluations. A context evaluation's main objectives are to

- Set boundaries around and describe the setting for the intended program or other improvement effort
- Identify intended beneficiaries and assess their needs
- Identify problems or barriers to meeting the assessed needs
- Identify relevant, accessible assets and funding opportunities that could be used to address the targeted needs
- Provide a basis for setting improvement-oriented goals
- Assess the clarity and appropriateness of improvement-oriented goals
- Provide a basis for judging outcomes of the subject program

Table 13.3 Four Types of Evaluation and Their Objectives, Methods, and Uses

	Types of Evaluation			
	Context Evaluation	**Input Evaluation**	**Process Evaluation**	**Product Evaluation**
Objectives	To define the relevant context, identify the target population and assess its needs, identify opportunities for addressing the needs, diagnose problems underlying the needs, and judge whether program goals and priorities are sufficiently and appropriately responsive to the assessed needs	To identify and assess system capabilities and alternative program strategies and then assess the chosen strategy's procedural design, budget, schedule, and staffing and stakeholder involvement plans	To identify or predict defects in the procedural design or its implementation, provide information for preprogrammed implementation decisions, affirm activities that are working well, and record and judge procedural events and activities	To identify intended and unintended outcomes; relate them to goals and assessed needs and to context, input, and process information; and judge accomplishments in terms of such factors as quality, worth, probity, equity, cost, safety, and significance
Methods	System analysis, surveys, document review, secondary data analysis, hearings, interviews, focus groups, diagnostic tests, case studies, site visits, epidemiological studies, and the Delphi technique	Document analysis, interviews, literature review, visits to exemplary programs, advocate teams studies, checklists, pilot tests, and content analysis	Monitoring the program's potential procedural barriers and remaining alert to unanticipated ones, obtaining information for implementation decisions, documenting the actual processes and costs, photographing progress, and regularly interacting with and reporting to staff and other stakeholders	Objective measurement, attitude scales, documentation of participation, interviews, photographic records, cost-effectiveness analysis, effect parameter analysis, goal-free evaluation, experimental design, time-series studies, surveys, content analysis, and significance tests
Uses	For deciding on the setting to be served, the goals associated with meeting needs or using opportunities, the priorities for budgeting time and resources, and the objectives associated with solving problems and planning needed program changes, and for providing a basis for judging outcomes	For selecting sources of support, solution strategies, and procedural designs (that is, for structuring, staffing, scheduling, and budgeting improvement activities), and for providing criteria for judging implementation	For implementing and refining the program design and procedures (that is, for effecting process and quality control), and for providing a log of the actual process and program costs for later use in interpreting outcomes	For deciding to continue, modify, or refocus a program, and for presenting a clear record of effects (intended and unintended, positive and negative), compared with assessed needs, targeted goals, and costs

Context evaluations may be initiated before, during, or even after a project, program, or other intervention. In the before case, organizations may carry out context evaluations as narrowly bounded studies to help set goals and priorities in a particular area. In the case of evaluations started during or after a program or other intervention, institutions often conduct and report on context evaluations in combination with input, process, and product evaluations. Here, context evaluations are useful for judging established goals and helping the audience assess the effort's worth in meeting beneficiaries' needs.

A context evaluation's methodology may involve collecting a variety of information about members of the target population and their surrounding environment and conducting various

types of analysis. A usual starting point is to ask the client and other stakeholders to help define the study's boundaries. Subsequently the evaluator may employ a variety of techniques to generate and test hypotheses about needed services or changes in existing services. These techniques might include reviewing documents, analyzing demographic and performance data, conducting hearings and community forums, conducting focus group sessions, and interviewing beneficiaries and other stakeholders.

The evaluator might construct a survey instrument to investigate identified hypotheses concerning the existence of beneficiaries' needs. Then he or she could administer it to a carefully defined sample of stakeholders. The evaluator could also make the survey instrument available more generally to anyone wishing to provide input, analyzing the two sets of responses separately.

The evaluator should examine existing records to identify performance patterns and background information on the target population. In a school district, information gathered might include immunization records; enrollment in different levels of courses; attendance; school grades; test scores; honors; graduation rates; participation in extracurricular activities; participation in special education; participation in free and reduced-fee meal programs; participation in further education; housing situations; employment and health histories; disciplinary records; or feedback from teachers, parents, former students, counselors, coaches, health personnel, librarians, custodians, administrators, or employers.

The evaluator might administer special diagnostic tests to members of the target population. He or she might engage an expert review panel to visit, closely observe, and identify needs, problems, assets, and opportunities in the targeted environment. The evaluator might conduct focus group meetings to review the gathered information, possibly using a consensus-building technique, such as Delphi, to solidify agreements about priority needs and goals. These procedures contribute to an in-depth perspective on the school district's functioning and highest-priority needs.

Often audiences need to view an effort within both its present setting and its historical context. Considering the relevant history helps decision makers avoid past mistakes. Thus, the methodology of context evaluation includes historical analysis and literature review as well as methods aimed at characterizing and understanding current environmental conditions. After the initial context evaluation, an organization often needs to continue collecting, organizing, filing, and reporting context evaluation data, because needs, problems, assets, and opportunities are subject to change.

In some situations, the evaluator should look beyond the local context to ascertain whether a program has widespread relevance. For example, a successful early childhood program might produce a ripple effect that eventually improves early childhood programming far beyond the program's setting. In such cases, the evaluator would judge the program not only on its worth in addressing the needs of targeted beneficiaries but also on its significance in serving beneficiaries outside the program's area of operation. When a context evaluation shows that a proposed program has widespread significance, the program developer can make an especially strong case for external financial support.

Context evaluations have a wide range of possible constructive uses. A context evaluation might provide a means by which an administrator can communicate with constituents to gain a shared conception of the organization's strengths and weaknesses, needs, assets, opportunities, and priority problems. A program developer could use context evaluation information to support a request for external grants or contracts. A university might use a context evaluation to convince a funding agency that it directed a proposed program at an urgent need or to convince a state legislature to increase the institution's funding. A social service organization might use context evaluation information to formulate objectives for staff development or to identify target populations for priority assistance. A school would use a context evaluation to help students and their parents or advisers focus their attention on developmental areas requiring more progress. An institution also could use context evaluation information to help decide how to make itself stronger by cutting marginally important or ineffective programs.

Context evaluation information is particularly useful when an organization needs to assess the worth and significance of what an intervention accomplished. Here the organization assesses whether the investment in improvement effectively addressed the targeted needs of intended beneficiaries. Also, evaluators refer to context evaluation findings to assess the appropriateness of goals and the relevance of program plans. Similarly, they use context evaluation findings to examine how an intervention's process is effecting improvements outside the local setting. Considering such uses, an organization can benefit greatly by establishing, keeping up to date, and using information from a context evaluation database.

Input Evaluation

An input evaluation's main orientation is toward helping prescribe a program approach by which to make needed changes. To this end, evaluators search out and critically examine potentially relevant approaches, including the one already being used. Input evaluation has a bearing on the success or failure and efficiency of a change effort. Initial decisions to allocate resources constrain change programs. A potentially effective solution to a problem will have no possibility of impact if a planning group does not at least identify it and assess its merits. A secondary orientation of an input evaluation is toward informing interested parties about what program approach was chosen, over what alternatives, and why. In this sense, input evaluation information is an important source of a developer's accountability for the design and budgeting of an improvement effort.

Essentially, an input evaluation should involve identifying and rating relevant approaches (including associated equipment and materials) and assist decision makers in preparing the chosen approach for execution. An evaluator should also search through the client's environment for political barriers, financial or legal constraints, and potentially available resources. The overall intent of an input evaluation is to help decision makers examine alternative program strategies for addressing assessed needs of targeted beneficiaries, evolve a workable program plan and appropriate budget, and develop an accountability record for defending the program's procedural and resource plans. Another important function is to help program leaders avoid the wasteful practice of pursuing proposed innovations that predictably would fail or at least waste resources.

Evaluators conduct input evaluations in several stages, which occur in no set sequence. An evaluator might first review the state of practice in meeting the specified needs and objectives. This process could include a number of possible components:

- Reviewing relevant literature
- Visiting exemplary programs
- Consulting experts and government representatives
- Querying pertinent information services (especially those on the World Wide Web)
- Reviewing a pertinent article in *Consumer Reports* or a similar publication that critically reviews available products and services
- Inviting proposals from involved staff

Evaluators might organize such information in a special planning room, possibly engaging a special study group to investigate it or conducting a special planning seminar to analyze the material. An evaluator would use the information to assess whether potentially acceptable solution strategies exist. He or she would rate promising approaches on relevant criteria, such as the following:

- Responsiveness to assessed needs of targeted beneficiaries
- Responsiveness to targeted problems in the organization
- Use of special funding programs or other relevant opportunities
- Potential effectiveness
- Cost
- Political viability
- Administrative feasibility
- Potential for important impacts outside the local area

Next, the evaluator could advise the decision makers about whether they should seek a novel solution. In seeking an innovation, the client and evaluator might document criteria the innovation should meet, structure a request for proposal, obtain competing proposals, and rate them on the chosen criteria. Subsequently the evaluator might rank the potentially acceptable proposals and suggest how the institution could combine their best features. The evaluator might also conduct a hearing to obtain additional information. He or she could ask staff and administrators to express concerns, so that together they might appraise resources and barriers that the institution should consider when installing the intervention. Members of the program planning group (that is, the client group) could then use the accumulated information to design and budget for what they see as the best combination strategy and action plan.

Input evaluations have several applications. A chief one is in preparing a proposal for submission to a funding organization or policy board. Another is in assessing one's existing

practice, whether or not it seems satisfactory, against what is being done elsewhere and what is proposed in the literature. Input evaluation has been used in the Dallas Independent School District; the Des Moines, Iowa, Public Schools; and the Shaker Heights, Ohio, School District. These districts used it to decide whether locally generated proposals for innovation were likely to be cost effective. The public school district for Detroit also used input evaluation to generate and assess alternative architectural designs for new school buildings. And the U.S. Marine Corps (USMC) used input evaluation to replace its system for evaluating officers and enlisted personnel; it did so by identifying and evaluating a wide range of personnel evaluation systems used by other military branches and in business and industry, by engaging planning teams to invent new creative approaches, by evaluating the capacity of all identified approaches to address the USMC's need for the new system, and ultimately by selecting and installing the preferred new approach. In addition to informing and facilitating decisions, input evaluation records help authorities defend their choice of one course of action over other possibilities. Administrators and policy boards can find input evaluation records useful when they must publicly defend sizable expenditures for new programs.

The advocate teams technique is a procedure designed specifically for conducting input evaluations. This technique is especially applicable in situations where an institution lacks effective means to meet targeted needs and stakeholders hold opposing views on what strategy the institution should adopt. Using this technique, the evaluator convenes two or more teams of experts and stakeholders, giving the teams the objectives, background data on needs, specifications for a solution strategy, and criteria for evaluating the teams' proposed strategies. The evaluator may staff these teams to match members' preferences and expertise to the nature of initial ideas about appropriate alternative strategies. Evaluators should do so especially if stakeholders severely disagree about what type of approach they would accept. Alternatively, if seeking creative, "out-of-the box," alternative strategies, the evaluator might staff the advocate teams with two or more groups of highly creative experts possessing significant knowledge and field experience in the general problem area. However the advocate teams are staffed, they compete in isolation from one another to develop a winning solution strategy. A panel of experts and stakeholders rates the advocate team reports. The institution might also field-test the teams' proposed strategies. Subsequently the institution would operationalize the winning strategy. Alternatively, it might combine and operationalize the best features of the two or more competing strategies.

The advocate teams technique's advantages are that it provides a systematic approach for

- Designing interventions to meet assessed needs
- Generating and assessing competing strategies
- Exploiting bias and competition in a constructive search for effective alternatives
- Addressing controversy and breaking down stalemates that stand in the way of progress
- Involving personnel from the adopting system in devising, assessing, and operationalizing improvement programs
- Documenting why a particular solution strategy was selected

Additional information, including a technical manual and the results of five field tests of the technique, is available in Reinhard (1972).

Process Evaluation

A process evaluation includes an ongoing check on a plan's implementation and documentation of the associated processes. One objective is to provide staff and managers with feedback about the extent to which they are carrying out planned activities on schedule, as planned and budgeted, and efficiently. Another is to guide staff to improve the procedural and budgetary plans appropriately. Typically staff cannot determine all aspects of such plans when a program starts. Also, they must alter the plans if some initial decisions are unsound or not feasible. Still another objective is to periodically assess the extent to which participants accept and can carry out their roles. In process evaluations, evaluators should contrast activities and expenditures with the plan and budget, describe implementation problems, and assess how well the staff has addressed them. They should document and analyze the effort's costs. Finally, they should report on how observers and participants judged the quality of the program's implementation.

The linchpin of a sound process evaluation is the process evaluator. More often than not, staff members' failure to obtain guidance for implementation and to document their activities and expenditures is due to a failure to assign anyone this work. Sponsors and institutions too often assume erroneously that the managers and staff will adequately evaluate program implementation as a normal part of their assignments. Managers and staff may routinely do some review and documentation through staff meetings, minutes of the meetings, and periodic accounting reports, for example, but these components do not fulfill the requirements of a sound process evaluation. Experience has shown that program directors can usually meet these requirements well only by assigning an evaluator to provide ongoing program review, feedback, and documentation.

A process evaluator has much work to do in monitoring and documenting an intervention's activities and expenditures. Initially, the process evaluator could review the relevant program strategy, work plans, the budget, and any prior background evaluation to identify what planned activities he or she should monitor. Possible examples of such activities are delivering services to beneficiaries, hiring and training staff, supervising staff, conducting staff meetings, monitoring and inspecting work flow, securing and maintaining equipment, ordering and distributing materials, controlling finances, documenting expenditures, managing program information, and keeping constituents informed.

Bearing in mind such process evaluation issues as those just mentioned, the process evaluator could develop a general schedule of data collection activities and begin carrying them out. Initially these probably should be as unobtrusive as possible so as not to threaten staff members, get in their way, or constrain or interfere with program implementation. As rapport develops, the process evaluator can use a more structured approach. At the outset, the process evaluator should obtain an overview of how the work is going. He or she could visit and observe centers of activity; review pertinent documents (especially the work plan, budget, accounting reports, and minutes of meetings); attend staff meetings; and interview

key participants. The evaluator then could prepare a brief report that summarizes the data collection plan, findings, and observed issues. He or she should highlight existing or impending implementation problems that the staff should address. The evaluator could then deliver this report at a staff meeting.

The process evaluator might invite the staff's director to lead a discussion of the report. The program team could then use the report for decision making as it sees fit. Also, the evaluator could review plans for further data collection and the creation of the subsequent report with the staff and ask them to react to the plan. Staff members could identify what information they would find most useful at the next meeting. They could also suggest how the evaluator could best collect certain items of information—for example, by using observations, staff-kept diaries, interviews, or questionnaires. The evaluator should ask staff members when they could best use the next evaluation report.

Using this feedback, the evaluator would schedule future feedback sessions. He or she would modify the data collection plan as appropriate and proceed accordingly. The evaluator should continually show that process evaluation helps staff members carry out their work through a kind of quality assurance and ongoing problem-solving process. He or she should also sustain the effort to document the program's implementation and lessons learned for use in the future summative evaluation report.

The evaluator should periodically report on how well the staff is carrying out the work plan and integrating it into the surrounding environment. He or she should describe main deviations from the plan, and should note variations in how different persons, groups, or sites are carrying out the plan. He or she should also characterize and assess the ongoing planning activity and record of expenditures.

Staff members use process evaluation information to guide activities, correct faulty plans, and maintain accountability records. Some managers use regularly scheduled process evaluation feedback sessions to keep staff members on their toes and abreast of their responsibilities. Process evaluation records are useful for accountability, because funding agencies, policy boards, and constituents typically want objective and substantive confirmation of whether grantees did what they had proposed and expended allocated funds appropriately. Process evaluations can also help external audiences learn what was done in an enterprise and at what cost, in case they want to conduct a similar effort. Such information is also useful to new staff members as part of their orientation to what has gone before. Moreover, process evaluation information is vital for interpreting product evaluation results. One needs to learn what was done in a program before deciding why program outcomes turned out as they did.

Over the years, the Evaluation Center at Western Michigan University has developed and employed a procedure labeled the "traveling observer technique" (for example, Alexander, 1974; Evers, 1980; Nowakowski, 1974). This technique most heavily addresses process evaluation data requirements but, like other techniques, also provides data of use in context, input, and product evaluation. In using the technique, the evaluator or head of the evaluation team selects and engages an investigator—such as a sociologist, anthropologist, or advanced evaluation graduate student—to carry out a well-defined traveling observer assignment. In such an assignment the traveling observer collects particular process information as determined by the evaluator

or head evaluator; gathers the assigned information, typically at several program field sites; and performs other assigned tasks, such as scheduling and arranging for a team of experts to visit program sites and assess the program's effectiveness. Basically, the traveling observer investigates and characterizes how the program's staff members are carrying out the program at the different program locations and then reports the findings to the evaluation's director and other evaluation team members, such as a site visit team of experts.

The traveling observer follows a set schedule of data collection and writes and delivers reports according to preestablished formats and reporting specifications. Before entering the field, the traveling observer develops a traveling observer handbook (Alexander, 1974; Nowakowski, 1974; Reed, 1989; Sandberg, 1986; Sumida, 1994). With the head evaluator, he or she tailors this evaluation tool to the evaluation's questions. The handbook includes the following parts:

- The traveling observer's credentials
- Evaluation questions
- A description of the field sites and program activities
- Program contact personnel and telephone numbers
- Maps showing program locations
- Data sources suggested, including interviewees and pertinent documents
- Protocols for contacting field personnel and obtaining needed permissions and cooperation
- Rules concerning professional behavior expected
- Safeguards to help the traveling observer avoid co-optation by program staff
- Sampling plans, including both preset samples and exploratory grapevine sampling
- Recommended data collection procedures
- Data collection instruments
- The data collection schedule
- The daily log or diary format
- Rules for processing information and keeping it secure
- The audience for traveling observer feedback
- The reporting specifications and schedule, including guidelines for interim progress reports, briefing sessions, and expense reports
- Criteria for judging traveling observer reports
- Rules about communicating and disseminating findings, including provisions for reporting to those who supplied data for the traveling observer study
- Any responsibilities for scheduling and facilitating follow-up investigations (for example, a site visit by a team of experts)
- Issues that may arise and what to do about them

- A form for the traveling observer's periodic self-assessment
- The budget to support the traveling observer's work, including spending limitations and reporting requirements

In an early application of this technique, the Evaluation Center sent out traveling observers as advance persons to do initial investigations of two $5 million statewide National Science Foundation programs (located in a state on the U.S. East Coast and a state on the U.S. West Coast). The Evaluation Center assigned the traveling observers to prepare the way for follow-up site visits by high-level teams composed of national experts in science, mathematics, technology, evaluation, and education. Each program included many projects at many sites across the state. The evaluation budget was insufficient to send the five-member teams of high-priced experts to all the potentially important sites, so the center preprogrammed and sent a traveling observer to study the program in each state. Each traveling observer spent two weeks investigating the program and prepared a report that contained findings and a tentative site visit agenda for the follow-up team of experts that would investigate the program. The traveling observers also contacted program personnel to prepare them for the follow-up visits and to ensure that they understood and supported the evaluation. On the first day of each team's site visit, the corresponding traveling observer distributed her or his report to the team and explained the results. The traveling observers also oriented the teams to the geography, politics, personalities, and other characteristics of each program. They presented each team with a tentative site visit agenda and answered team members' questions. The traveling observers' recommended plans for the site visit teams included sending different members of the team to different program sites and hosting some whole-team meetings with key program personnel.

During the week-long site visits, the traveling observers remained accessible by telephone so they could help the site visit team members. At the end of this study, the center engaged Michael Scriven to evaluate the evaluation. He reported that the traveling observer reports were so informative that except for taking into account the credibility added by the national experts, the traveling observers could have evaluated the programs successfully without the experts. Overall, the Evaluation Center has found that the traveling observer technique is a powerful evaluation tool; it is systematic, flexible, efficient, and inexpensive. Its focal use is to help evaluate a program's implementation, but it sets traveling observer study in the context of assessed needs, program structure, and outcomes. It also is useful in preparing for follow-up, in-depth site visits.

Product Evaluation

The purpose of a product evaluation is to measure, interpret, and judge an enterprise's outcomes. Its main objective is to ascertain the extent to which the evaluand met the needs of all the rightful beneficiaries. Feedback about outcomes is important both during an activity cycle and at its conclusion. Product evaluators should assess intended and unintended outcomes and positive and negative outcomes. Moreover, they often should extend product evaluation to assess long-term outcomes.

In conducting a product evaluation, the evaluator should gather and analyze stakeholders' judgments of the program. Sometimes the product evaluation should include a comparison of the effort's outcomes with those of similar enterprises. Frequently clients want to know whether a program achieved its objectives and was worth the investment. If appropriate, evaluators should interpret whether poor implementation of the work plan caused poor outcomes. Finally, a product evaluation should usually view outcomes from several vantage points: in the aggregate, for subgroups, and sometimes for individuals.

Product evaluations follow no set algorithm, and many methods are applicable. Evaluators should use a combination of techniques. This aids them in making a comprehensive search for outcomes. It also helps them cross-check the various findings.

To assess performance beyond objectives, evaluators need to search for unanticipated outcomes, both positive and negative. They might conduct hearings or group interviews to generate hypotheses about the full range of outcomes and follow these up with clinical investigations intended to confirm or disconfirm the hypotheses. They might conduct case studies of the experiences of a carefully selected sample of participants to obtain an in-depth view of the program's effects. They might survey, by telephone or mail, a sample of participants to obtain their judgments of the program and their views of both positive and negative findings. They might ask participants to submit concrete examples of how the program influenced their work or well-being. These could be written pieces, other work products, or a new job status. They might engage observers to identify program and comparison groups' achievements. They might also compare identified program achievements against a comprehensive checklist of outcomes of similar programs. Also, they might compare recently assessed outcomes with outcomes identified at one or more prior points in time. Trend analysis can be invaluable when outcome variables have been measured repeatedly (for example, prior to, during, and after a period of intervention).

An evaluator might conduct a goal-free evaluation (Scriven, 1991), whereby he or she engages an investigator to find whatever effects an intervention is producing or has produced, not just those associated with the program's goals. The evaluator informs the goal-free investigator of the identity of the targeted beneficiaries and the environment in which the program operates, but purposely prevents the goal-free investigator from learning the program's goals. The point is to keep the goal-free investigator from developing tunnel vision focused on stated goals. The goal-free investigator searches through the program's environment to identify what the program is actually doing and what outcomes are evident. He or she also assesses the needs of the program's targeted beneficiaries. In analyzing the findings of the goal-free evaluation, the evaluator compares the identified outcomes to the assessed needs of beneficiaries and draws conclusions about the program's effectiveness. The goal-free evaluation technique provides evaluators and clients with a unique approach to assessing an intervention's value, whatever its goals. The technique is especially powerful for uncovering a program's side effects. So long as the lead evaluator keeps the goal-free study separate from other parts of the involved program evaluation, it is permissible and often useful simultaneously to conduct both goal-free and goals-based evaluations and subsequently to compare their results.

Reporting of product evaluation findings may occur at different stages. Evaluators may submit interim reports during each program cycle. These should show the extent to which the intervention is addressing and meeting targeted needs. End-of-cycle reports may sum up the results achieved. Such reports should offer an interpretation of the results in light of assessed needs, costs incurred, and execution of the plan. Evaluators may also submit follow-up reports to assess long-term outcomes. In such reports, evaluators might provide analysis of the results in the aggregate, for subgroups, and for individuals.

People use product evaluations to decide whether a given program, project, service, or other enterprise is worth continuing, repeating, or extending to other settings. A product evaluation also should provide direction for modifying the enterprise or replacing it so that the institution will more cost-effectively serve the needs of all targeted beneficiaries. Of course, it should help potential adopters decide whether the approach merits their serious consideration. Product evaluations have psychological implications, because by showing signs of growth or superiority to competing approaches, they reinforce the efforts of both staff and program recipients; or they may dampen enthusiasm and reduce motivation when the results are poor.

In regard to the latter point, evaluators should not publicly release product evaluation findings too soon. A program requires time to achieve results for which it should be held accountable. Premature release of a product evaluation report might unjustly discourage continuation of the program because no positive results were found. If a public report containing product evaluation findings is delayed for a reasonable amount of time, the evaluator might discover late-blooming, important outcomes that would support continuation of the program. Also, an evaluator can stifle program staff members' creativity by being overzealous in conducting and reporting on a product evaluation during a program's exploratory stage. Of course, the evaluator can respond appropriately to staff requests for ongoing formative product evaluation findings; usually such early interim results should be shared only with the program's staff members as an aid to their quest for success. Rules of thumb are that evaluators should be low key in conducting a product evaluation early in a program and should not report product evaluation findings to anyone beyond the program staff until they have had ample time to install and stabilize procedural plans. The evaluator should distribute product evaluation findings to right-to-know audiences after the program has had a fair chance to mature and produce its outcomes. Clearly, evaluators need to exercise professional judgment and discretion in deciding matters of conducting and reporting product evaluation findings.

Finally, product evaluation information is an essential component of an accountability report. When authorities document significant achievements, they can better convince community and funding organizations to provide additional financial and political support. If authorities learn that the intervention made no important gains, they can cancel the investment. This frees up funds for worthy interventions. Moreover, other developers can use the product evaluation report to help decide whether it is appropriate to pursue a similar course of action.

Use of the CIPP Model as a Systems Strategy for Improvement

The CIPP model is a social systems approach to evaluation. A social system is an interrelated set of activities that function together to fulfill a mission and achieve defined goals within a certain context. Within this view, evaluation appropriately promotes and assists with goal achievement and ongoing program improvement. The model opposes the view that evaluations should typically be one-shot investigations, should be conducted solely by evaluators, or should be merely instruments of accountability for externally funded programs. Instead, it treats evaluation as a tool by which evaluators, in concert with stakeholders, can help programs, projects, and other services perform better for the beneficiaries. Fundamentally the model is designed to promote growth. Optimally, a CIPP evaluation is a sustained, ongoing effort to help an organization's leaders and staff obtain, organize, and use feedback systematically to validate goals, meet the targeted needs of beneficiaries, and pass accountability examinations.

The flowchart in Figure 13.2 displays the CIPP model's systems orientation. To explain and illustrate this flowchart, we will discuss evaluation in the hypothetical context of a social ministries committee's efforts to meet the needs of foster care youths in its geographical area. The committee includes representatives from area churches and has a long-standing, positive record of providing foster care children and families with support. Recently, however, this committee conducted a context evaluation to determine whether certain needs of foster care children were being met. This context evaluation activity in general is represented in the upper left-hand corner of Figure 13.2.

On reviewing a published national report on older foster care children, the committee became concerned that area eighteen- and nineteen-year-old youths probably were vulnerable to serious culture shock when they left the area's foster care system. Study of such youths in the local area uncovered startling facts. Over the five years after leaving their foster care families, about 40 percent of the area's foster teenagers soon became homeless, suffered emotional breakdowns, or committed crimes and were incarcerated. The committee found further that these older youths often needed, but did not receive, assistance in securing shelter, clothing, food, employment, further education, health care, and psychological and emotional support. A main reason for such unmet needs was that once a foster youth turns eighteen, foster parents receive no further financial assistance. Consequently, many of them are unable or unwilling to continue housing the youth. Subsequently, many eighteen-year-old foster care youths are abruptly turned out of their foster care home. They are not ready to fend for themselves and encounter serious difficulties. Regrettably, they have nowhere to belong. Based on these context evaluation findings, the committee decided to mount a program for the area's older foster youths called Transition from Foster Care to Productive Adult Life.

In planning the program, the committee consulted local community agencies and service groups to ascertain whether they already possessed some appropriate and available comprehensive support strategy that the church group could get funded and then immediately install. They found no such strategy. Therefore, with the assistance of a volunteer evaluation specialist, the church committee proceeded to plan and conduct an input evaluation aimed at identifying

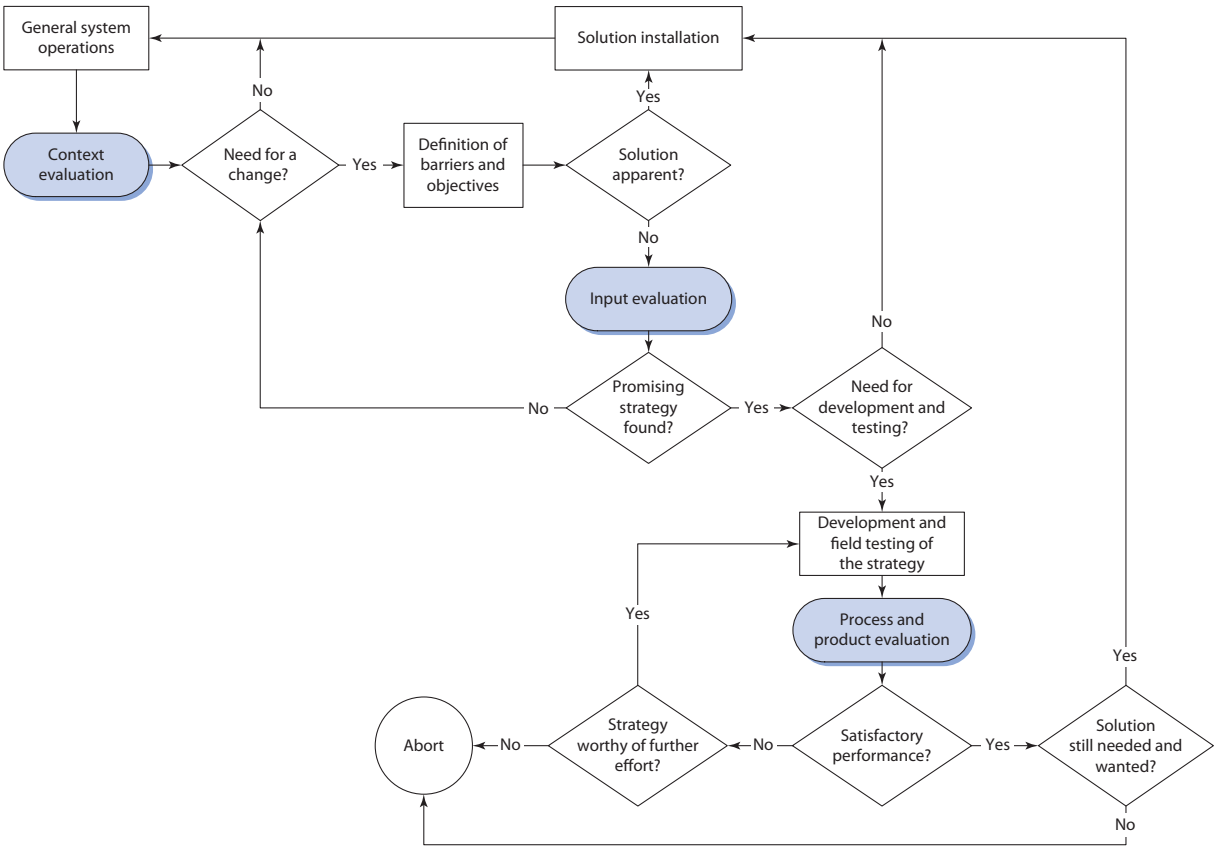

Figure 13.2 Flowchart of a CIPP Evaluation in Fostering and Assessing System Improvement

and rating the relevance and feasibility of alternative ways to serve youths who are exiting the foster care system. The evaluator emphasized that development of competing proposals would stimulate creativity and identify an appropriately broad range of problem-solving strategies.

This input evaluation began with a review of the relevant literature and queries to state and local support organizations with some experience in managing or assisting with foster care. The committee conducted focus groups, meetings at area organizations, interviews with a wide range of community personnel, and a community-wide conference to investigate the issue. Those providing information and deliberating were from the local foster care agency, Habitat for Humanity, area courts, law enforcement, city management, churches, a community foundation, the Salvation Army, a local hospital, and a university. These sources also included former foster care parents and children and military recruiters. The respondents were asked to give their perceptions of the needs and problems of this group of youths and to identify ways they thought area groups could respond effectively.

The volunteer evaluator next organized the social ministries committee into two proposal-writing groups and supplemented each group with additional volunteers from the community. Each group was given the assignment of using the information obtained so far to write a plan for addressing the identified needs of the subject foster care youth. Criteria for developing and evaluating the proposals were drawn from the needs and problems identified in the context evaluation and from criteria for funding proposals from prospective funding organizations. In general, these criteria addressed such topics as responsiveness to the defined needs and problems of older foster care youths; goals for the foster care intervention; consistency with relevant research on children and youth; compatibility with the community's existing foster care support system; availability of committed volunteers; cost of development; long-term affordability; acceptability to foster care youths; the potential benefits to the local community and economy of helping foster care youth become productive, self-actualized citizens; inclusion of formative and summative evaluation plans; and responsiveness to any unique proposal requirements of prospective funding organizations. Each team was also given an outline of points to be included in its proposal.

On completion of the proposals, the evaluator convened and chaired a group of area persons concerned and knowledgeable about the foster care issue. This group evaluated the alternative proposals against the prescribed criteria, identified the strongest parts of each proposal, and recommended how these might be merged into an overall plan. The group made recommendations about how the merged plan could be used to develop different proposals for different funding organizations.

Using the results of this input evaluation, the committee developed the overall program plan. It included a resource center where youths could obtain clothing, bedding, kitchen utensils, and other housewares; a program to recruit, train, and engage community members to serve as mentors to the youths; a committee of local business representatives to help youths find jobs; a committee of local health care professionals to help youths receive needed health care; a scholarship program to help qualified youths pursue further education; a support group with regular meetings at which youths could share and address their problems and develop life skills; screened and approved host families to rent rooms to youths; and the development and ongoing support of several supervised independent-living group homes.

The committee subsequently contacted area churches, community service groups, colleges and universities, Habitat for Humanity, Big Brothers–Big Sisters, hospitals, local media, and several prospective funding organizations for financial and other kinds of assistance. The committee informed these parties about the context evaluation and input evaluation findings and summarized the resulting program plan. Following discussions with these groups, the committee wrote and submitted specific funding proposals and requests for assistance in keeping with the overall plan and the parts of the work that the different funding and other groups found to be within their targeted areas of support or involvement. Subsequently, the social ministries committee received several grants and other kinds of support and proceeded to oversee and help implement the overall program and each of its parts.

Fortunately, the volunteer evaluator had convinced the committee to build continuing evaluation into its plans. Accordingly, the evaluator continued to support the committee's work by coordinating and assisting with both process and product evaluations. The legwork in these evaluations was done by community volunteers and university graduate students, who with the volunteer evaluator became the effort's evaluation team. At monthly meetings, the evaluation team presented the committee with process evaluation reports focused on how well each part of the Transition from Foster Care to Productive Adult Life program was being carried out. Periodically they presented product evaluation findings that focused on the overall program's successes and failures with individual youths and groups of youths and on the success and cost-effectiveness of each part of the program. In particular, they compared the identified outcomes to the needs and problems found in the original context evaluation. The committee used the feedback to gauge the success of the program and each component, solve emergent problems, adjust plans, carry on with the work, write new proposals, and seek additional funding and other forms of support.

At appropriate intervals, the committee compiled evaluation results and presented them to its support groups. Key issues addressed were whether the program and its individual components were succeeding and whether its long-term implementation was warranted and sustainable. The results proved to be positive and indicative of the program's viability for the long term. All the support organizations were convinced of the program's value and helped the committee establish an ongoing, stable base of monetary and nonmonetary support. Based on the valuable contribution of evaluation to this effort, the committee wisely decided to continue its context evaluation surveillance of the needs of older foster care youth, to conduct additional input evaluations as needed and the associated proposal writing, and also to continue the process and product evaluations of the installed interventions. Clearly, credit for this improvement effort's success belongs to the charitable, sustained work of the foster care volunteers; however, it is also clear that systematic evaluation provided them with an invaluable aid to effectively identifying and addressing the needs of the foster care youth.

Summary

CIPP is an acronym that denotes the CIPP model's four core types of evaluation. In context evaluations, evaluators assess needs, problems, opportunities, and assets as bases for setting and judging goals and priorities. In input evaluations, evaluators identify and assess alternative

approaches and plans for meeting assessed needs and achieving defined goals. In process evaluations, evaluators monitor, document, and provide feedback for strengthening program implementation. In product evaluations, evaluators identify and assess both intended and unintended outcomes—positive as well as negative. The model is configured for both formative use in developing and conducting programs and summative use in judging completed programs and meeting accountability requirements.

The CIPP model originally was developed because school districts across the nation needed to effectively address evaluation and accountability requirements of federal programs for reforming education and because such traditional methods as objectives-based evaluation, experimental design, and standardized achievement testing proved inappropriate to the evaluation task. In general, the CIPP model is configured to enable and guide comprehensive, systematic examination of social and educational programs that occur in the dynamic, septic conditions of the real world. The model's main orientation is toward helping those who sponsor and conduct programs to obtain and use systematic evaluative feedback as an effective aid to continual program improvement. Since its creation, the model has been further developed and adapted; it has been applied in virtually all disciplines and service areas and across the world; and it has been applied to evaluations of a wide range of objects beyond programs, including organizations, personnel, equipment, materials, policies, and evaluations.

The CIPP model defines evaluation, generally, as the systematic investigation of some object's value and, operationally, as the process of delineating, obtaining, reporting, and applying descriptive and judgmental information about an object's value, as defined by such criteria as quality, worth, probity, equity, feasibility, cost, efficiency, safety, and significance.

The model is grounded in an objectivist quest for clear, unambiguous answers. It subscribes to the values of a free, democratic society. It stresses that evaluation's most important purpose is not only to prove but to improve. It provides for equitable, meaningful engagement of stakeholders in the evaluation process. It offers a template for organizations to use in institutionalizing and mainstreaming systematic evaluation. It employs a wide range of quantitative and qualitative methods—including archival investigation, logic models, checklists, rating scales, interviews, questionnaires, standardized tests, advocate teams, traveling and resident observers, photographic records, focus groups, time-series studies, experimental design, cost-effectiveness studies, content analysis, effect parameter analysis, significance tests, and goal-free evaluation—and calls for triangulation of information. Fundamentally, the model requires that evaluations be guided and assessed against the professionally defined standards of utility, feasibility, propriety, accuracy, and evaluation accountability.

The model portrays evaluation as essential to societal progress and the well-being of individuals and groups. It embodies the contention that societal groups cannot make their programs, services, and products better unless they learn where these are weak and strong. They cannot convince consumers to buy or support their services and products unless their claims about the value of these services are valid and honestly reported. Institutional personnel cannot meet all of their institution's evaluation needs if they do not both contract for external evaluations and also build and apply capacity to conduct internal evaluations. Evaluators

cannot defend their evaluative conclusions unless they key them to both sound information and clear, defensible values. Moreover, internal and external evaluators cannot maintain credibility for their evaluations if they do not subject them to metaevaluations against appropriate standards. The CIPP model is supported by an extensive theoretical and pragmatic literature.

REVIEW QUESTIONS

1. Why and how was the CIPP model developed?
2. What are the similarities and differences between the concepts of needs assessment and context evaluation?
3. What is the essential meaning of input evaluation, and what are at least two illustrations of its use?
4. What is the relationship between the concepts of context, input, process, and product evaluations and the concepts of formative and summative evaluations?
5. What are examples of the use of process evaluations for formative and summative purposes?
6. What is the traveling observer technique, how is it applied within the framework of the CIPP model, and what are advantages of using the approach?
7. What is meant by the CIPP model's objectivist orientation?
8. What is the role of values in the CIPP model, and how are these identified and applied?
9. What is the advocate teams technique, and what is an illustration of its use in an input evaluation?
10. What is meant by the claim that the CIPP model is a social systems approach to evaluation, and what is the value of this orientation?

Group Exercises

Exercise 1

The CIPP model operationally defines evaluation as the process of delineating, obtaining, reporting, and applying descriptive and judgmental information about an object's value, as defined by such criteria as quality, worth, probity, equity, feasibility, cost, efficiency, safety, and significance. All of this is done to guide decision making; support accountability; disseminate effective practices; increase understanding of the involved phenomena; and, as appropriate, make decision makers, stakeholders, and consumers aware of evaluands that proved unworthy of further use. How does implementation of this definition and description satisfy the 2011 Joint Committee standards' requirement that evaluations meet conditions of utility, feasibility, propriety, accuracy, and evaluation accountability?

Exercise 2

Suppose you were asked to conduct a product evaluation of a state's long-standing policy to prohibit smoking in public buildings. What might you look at within the product evaluation's subparts of reach to the targeted beneficiaries, effectiveness, sustainability, and transportability? Identify the different techniques you would choose to employ in the four subparts of the evaluation. Then give a justification for using each of the techniques.

Exercise 3

Suppose you have agreed to serve as a traveling observer in a program evaluation. Your assignment basically is to conduct a process evaluation of a university's two-week summer camp for eleventh- and twelfth-grade cheerleaders from ten area high schools. Before beginning the traveling observer assignment, you will need to develop the traveling observer handbook you will use to guide your evaluation fieldwork. To develop the handbook, you must first get the evaluator who selected you to clarify your traveling observer assignment. List the questions pertaining to clarification of your traveling observer assignment that you would address to the evaluator who engaged you.

Exercise 4

Using the same evaluation example as that in exercise 3, do the following:

- Make a case for employing goal-free evaluation methodology to help conduct a product evaluation of the cheerleading summer camp.
- List criteria for selecting an evaluator to conduct the goal-free evaluation.
- Write specifications for conducting the goal-free evaluation.
- Discuss whether the product evaluation could appropriately include a goals-based evaluation as well as the goal-free study, and if so, under what circumstances.

Suggested Supplemental Readings

Alkin, M. C. (Ed.). (2013). *Evaluation roots: A wider perspective of theorists' views and influences* (2nd ed.). Thousand Oaks, CA: Sage.

Candoli, I. C., Cullen, K., & Stufflebeam, D. L. (1997). *Superintendent performance evaluation: Current practice and directions for improvement*. Norwell, MA: Kluwer.

Evers, J. (1980). *A field study of goal-based and goal-free evaluation techniques*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Joint Committee on Standards for Educational Evaluation. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Nevo, D. (1974). *Evaluation priorities of students, teachers, and principals*. Unpublished doctoral dissertation, Ohio State University, Columbus.

Nicholson, T. (1989). Using the CIPP model to evaluate reading instruction. *Journal of Reading*, *32*, 312–318.

Reinhard, D. (1972). *Methodology development for input evaluation using advocate and design teams*. Unpublished doctoral dissertation, Ohio State University, Columbus.

Stufflebeam, D. L. (1966, January). *Evaluation under Title I of the Elementary and Secondary Education Act of 1967*. Address delivered at the Title I Evaluation Conference sponsored by the Michigan State Department of Education, Lansing.

Stufflebeam, D. L. (1967). The use and abuse of evaluation in Title III. *Theory into Practice*, *6*, 126–133.

Stufflebeam, D. L. (1969). Evaluation as enlightenment for decision making. In A. Walcott (Ed.), *Improving educational assessment and an inventory of measures of affective behavior* (pp. 41–73). Washington, DC: Association for Supervision and Curriculum Development.

Stufflebeam, D. L. (1971). The use of experimental design in educational evaluation. *Journal of Educational Measurement*, *8*, 267–274.

Stufflebeam, D. L. (1985). Stufflebeam's improvement-oriented evaluation. In D. L. Stufflebeam & A. J. Shinkfield, *Systematic evaluation: A self-instructional guide to theory and practice* (pp. 151–207). Norwell, MA: Kluwer.

Stufflebeam, D. L. (2003). The CIPP model for evaluation. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 31–62). Norwell, MA: Kluwer.

Stufflebeam, D. L. (2003). Institutionalizing evaluation in schools. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 775–806). Norwell, MA: Kluwer.

Stufflebeam, D. L. (2004). The 21st century CIPP model: Origins, development, and use. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 245–266). Thousand Oaks, CA: Sage.

Stufflebeam, D. L. (2005). CIPP model (context, input, process, product). In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 60–65). 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 in education*. Itasca, IL: Peacock.

Stufflebeam, D. L., & Webster, W. J. (1988). Evaluation as an administrative function. In N. Boyan (Ed.), *Handbook of research on educational administration* (pp. 569–601). White Plains, NY: Longman.

Webster, W. J. (1975, March). *The organization and functions of research evaluation in a large urban school district*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Zhang, G., Zeller, N., Griffith, R., Metcalf, D., Shea, C., Williams, J., & Misulis, K. (2010, October). *Using the CIPP model as a comprehensive framework to guide the planning, implementation, and assessment of service-learning programs*. Paper presented at the annual meeting of the National Evaluation Institute, Williamsburg, VA.

CHAPTER 14

MICHAEL SCRIVEN'S CONSUMER-ORIENTED APPROACH TO EVALUATION

Overview of Scriven's Contributions to Evaluation

Scriven has sharply criticized several widely endorsed evaluation ideologies, including one that focuses on achieving the developer's objectives rather than meeting consumers' needs and one that portrays evaluation as value-free social science. He has proposed a rich array of concepts and methods designed to move evaluation from its objectives-based orientation to one keyed to assessed needs and societal ideals.

Moreover, he has characterized evaluation as a vital transdiscipline (that is, a discipline that provides other disciplines with services or tools) that inheres in all disciplined intellectual and practical endeavors and as one that needs to be developed and maintained as a discipline in its own right. He has called on evaluation theorists, educators, and practitioners to take necessary steps to advance their field in all of its important dimensions so that it can be applied meaningfully, competently, and systematically across the full range of societal enterprises.

Consistent with this interest, Scriven previously directed Western Michigan University's Interdisciplinary PhD in Evaluation (IDPE) program, which was designed by Daniel Stufflebeam in 2002, was directed by E. Jane Davidson from 2003 to 2004, was directed by Scriven from 2004 through 2007, and currently is being directed by Chris Coryn. The IDPE program is a collaborative effort between the Colleges of Arts and Sciences, Education and Human Development, Engineering and Applied Sciences, Health and Human Services, and the Evaluation

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **Michael Scriven's basic orientation to and definition of evaluation**
- **Scriven's critical appraisal of objectives-based and formative approaches to evaluation**
- **Scriven's 1983 attack on prevailing evaluation ideologies and associated methodological suggestions**
- **Key evaluation concepts of the consumer-oriented approach to evaluation, including formative and summative roles, needs assessment, goal-free evaluation, ascriptive evaluation, and metaevaluation**
- **Scriven's breakout of evaluation's basic acts into scoring, ranking, grading, and apportioning**
- **Scriven's recommended evaluation tools, including his Key Evaluation Checklist**
- **Scriven's recommended avenues to making defensible causal inferences**
- **Scriven's argument that evaluation is a self-referent discipline requiring all evaluators to self-assess and also obtain evaluations of their work**
- **Scriven's view of the three revolutions in evaluation**

Center to engage professors and recruit students from the full range of disciplinary backgrounds. The program prepares students to apply evaluation theory, methods, and practices across diverse disciplinary and service areas and to contribute to the development of evaluation as a recognized, respected transdiscipline (also see Coryn, Stufflebeam, Davidson, & Scriven, 2010). Along these lines, Scriven, along with Davidson, Coryn, and Daniela Schröter, founded the open-access *Journal of MultiDisciplinary Evaluation* (JMDE) in 2004—available at www.jmde.com—which is jointly sponsored by the IDPE program and the Evaluation Center. Since publication of its first issue, JMDE has attracted nearly six thousand subscribers in more than one hundred countries throughout the world, and papers published in the journal have been cited widely.

Scriven has many conceptual contributions to his credit (Scriven, 1974, 1991, 1993)—for example, formative evaluation, summative evaluation, ascriptive evaluation, metaevaluation, and goal-free evaluation. The most prominent of these are the concepts of formative and summative evaluation.

Summative evaluation enables consumers to decide whether a developed product or service—refined by the use of the evaluation process in its first, formative role—represents a sufficiently significant advance over the available alternatives to justify its purchase and use. Formative evaluation is mainly a proactive process for assessing and guiding the clarification of goals and implementation of plans.

Scriven (2007) has identified the key methods of evaluation as scoring, ranking, grading, and apportioning, and has noted that the logic of evaluation involves gathering and summarizing facts; collecting, clarifying, and verifying relevant values and standards; and synthesizing evidence and values into evaluative conclusions (also see Chapter 2). Although he sees experimental design as a valuable evaluation tool, he notes that this is only one of a range of methods for reaching defensible conclusions about cause and effect (Scriven, 2005a, 2009a). He has developed and continues to refine a practical tool—the Key Evaluation Checklist (KEC)—for applying his unique evaluation approach; the checklist can be retrieved from www.wmich.edu/evalctr/checklists. (Davidson [2005] has distilled Scriven's KEC into a pragmatic, step-by-step methodology.)

Beyond conceptual and methodological developments, since the 1970s Scriven has exerted strong, visionary leadership in helping to establish, organize, and administer professional evaluation organizations. These include the Evaluation Network (a professional society established in the late 1960s [see Chapter 1] and dedicated to improving training in evaluation) and subsequently, in 1986, the broader-gauge organization into which the Evaluation Network and Evaluation Research Society eventually evolved: the American Evaluation Association (AEA). In 2010 he established the Faster Forward Fund (3^F) aimed at financing and fostering efforts to accelerate needed theoretical and methodological developments in the evaluation discipline (Scriven, 2011d). As explained in this book's concluding section, in providing an orientation for 3^F, Scriven (2011a) reviewed past and in-progress revolutions in evaluation and pointed the way to a third needed revolution.

Clearly, as the thumbnail sketches just given show, for over forty years Scriven has been one of the evaluation field's most creative, productive, and influential iconoclasts (Heberger, Christie, & Alkin, 2010; Stufflebeam, 2013).

Scriven's Background

Scriven is a philosopher of science and an expert in critical thinking. He was born in England and raised in Australia. He earned his bachelor's degree in mathematics and his master's degree in applied mathematics and symbolic logic at the University of Melbourne. Subsequently, he completed his PhD in philosophy of science at Oxford University. He has served at ten universities in Australia, New Zealand, and the United States. From 2004 to 2007 he was professor of philosophy and associate director of the Evaluation Center at Western Michigan University as well as director of the IDPE program. Currently he holds a professorial appointment at Claremont Graduate University. The foundations of Scriven's evaluation approach are described in the book *Evaluation Roots* (Scriven, 2004b, 2013).

Scriven's Basic Orientation to Evaluation

We chose to refer to Scriven's approach as "consumer-oriented" evaluation so as to characterize his basic pragmatic orientation to addressing consumers' needs through evaluation. In an audiotape prepared for the American Educational Research Association (1969a), he stated that the proper role of the evaluator is that of "an enlightened surrogate consumer." In this role, the evaluator serves as an informed social conscience. Such service, he said, is the "foundation stone of professional ethics in evaluation work." Accordingly, evaluators armed with skills in obtaining pertinent and accurate information and with a deeply reasoned view of ethics and the common good should help professionals produce products and services that are of high quality, are best buys for consumers, and are in service to humankind. They also should help consumers identify and assess the merit, worth, and wide-ranging significance of alternative goods and services.

Scriven's practical approach to evaluation in general calls for identifying and ranking alternative programs and products that are available to consumers, based on the options' relative costs and effects and in consideration of the assessed needs of consumers and the broader society. He often has identified the magazine *Consumer Reports* as exemplary of what professional evaluation should contribute, although in his piece on product evaluation discussed later in this chapter (Scriven, 1994d), he criticized Consumers Union (publisher of *Consumer Reports*) for what he saw as a lowering of its technical standards and relaxation of its independence from commercial interests.

Scriven's Definition of Evaluation

According to Scriven (1991), "Evaluation is the process of determining the merit, worth and value of things, and evaluations are the products of that process" (p. 1). He has emphasized that evaluators must be able to arrive at defensible value judgments—or evaluative conclusions—rather than simply to measure things or determine whether goals have been achieved. Instead of accepting a developer's goals as givens, an evaluator, according to Scriven, must judge whether achievement of preordinate goals would contribute to the welfare of consumers. Regardless of goals, an evaluator must identify outcomes and assess their value

from the perspective of consumers' needs. Scriven (1973, 1974) advanced this position when he introduced the concept of goal-free evaluation, which entails a search for all effects of a program, irrespective of its goals.

Over the years, Scriven's suggested definition of evaluation has evolved, but its basic message has remained the same. In his classic 1967 article, "The Methodology of Evaluation," he defined evaluation as a methodological activity that

> consists simply in the gathering and combining of performance data with a weighted set of goal scales to yield either comparative or numerical ratings, and in the justification of (1) the data-gathering instruments, (2) the weighting, and (3) the selection of goals. (p. 40)

Although Scriven (1993) has urged evaluators to systematically determine the worth or merit of something, more recently he has added significance to these bottom-line criteria, stating that "one of the most important questions professional evaluators should regularly consider is the extent to which evaluation has made a contribution to the welfare of humankind, and, more generally, to the welfare of the planet we inhabit" (2004b, p. 183).

In discussing with Scriven the thrust of his definition of evaluation, we often have heard him say that evaluations are best executed by engaging an independent evaluator to render a judgment of some object based on the accumulated evidence about how it compares with competing objects in meeting the assessed needs of consumers. According to this view, evaluation is preferably comparative; by implication, it involves looking at comparative costs as well as benefits, and at how best to meet the needs of consumers; optimally, it is a professional activity involving systematic procedures; it should be conducted as objectively as possible and often by an independent evaluator; and it must culminate in judgments. We also note, however, that Scriven sees evaluation as a self-referent activity in which evaluators must evaluate their own work as well as obtain independent assessments of their evaluations (Scriven, 1991, 1993, 2007, 2009b).

Critique of Other Persuasions

Scriven has sharply criticized other views of evaluation and has used his critical analysis to extend his own position. He has charged that the Tylerian tradition (Scriven, 1991, 1993; Tyler, 1942), which sees evaluation as determining whether objectives have been achieved, is fundamentally flawed in that it is essentially value-free (meaning that the evaluator rather uncritically accepts the developer's values as reflected in stated goals and, based on unvalidated goals, makes claims or conclusions about something's quality or value that are not objectively defensible). He has argued that this approach is potentially invalid, because a developer's goals may be immoral, unrealistic, unrepresentative of the assessed needs of consumers, mainly in the developer's interest, or too narrow to encompass possibly crucial side effects. Instead of using goals to guide and judge effects, according to Scriven, evaluators should judge goals and not be constrained by them in the search for outcomes. Whether or not a program has been guided by meritorious goals, he believes, evaluators should search out all

of the results of a program (direct, indirect, intended, unintended, positive, and negative); assess the needs of consumers; and use both sets of assessments (that is, program results and consumer needs) to arrive at conclusions about the merit, worth, and significance of the program.

Scriven also took issue with the advice offered by Cronbach (1963), who criticized the prevalent practice of evaluating educational programs by using norm-referenced tests to compare the performance of experimental and control groups and had counseled the use of a more developmentally oriented approach. Cronbach advised against exclusive use of comparative experimental designs, suggesting that a variety of measures should be used to study a particular program while it is being developed and that the results should be used to help guide the program's development. In analyzing achievement test data, Cronbach preferred item analysis to help diagnose teaching and learning deficiencies to the more customary norms-based analysis of total test scores. Scriven argued that this advice by Cronbach clouded the important distinction between the goal and roles of evaluation and, in fact, tended to equate evaluation with only one of its roles—that is, the formative one. Building on this critique, Scriven extended his view of evaluation in his 1967 article, in which he introduced the terms *formative evaluation* and *summative evaluation*. Cronbach (1963) clearly played an important part in identifying the concepts of formative and summative evaluation, to which Scriven applied labels that have stood the test of time.

Formative and Summative Evaluation

In his 1967 article, Scriven argued that the evaluator's main responsibility is to make informed judgments. He emphasized that the goal of evaluation is always the same: to judge value. But, he continued, the roles of evaluation are enormously varied. They may "form part of a teacher-training activity, of the process of curriculum development, of a field experiment connected with the improvement of learning theory, [or] of an investigation preliminary to a decision about the purchase or rejection of materials" (pp. 40–41). He reasoned that the failure to distinguish between the goal of evaluation (to judge the value of something) and its roles (corresponding to constructive uses of evaluative information) has led to the dilution of what is called "evaluation" so that it no longer achieves its goal of assessing value. In other words, he said, evaluators, in trying to help improve programs, too often become co-opted and fail to judge the quality, value, and/or significance of programs. For Scriven, evaluation must provide an objective assessment of value.

With the paramount importance of the goal of evaluation firmly established, Scriven (1991, 1993, 1996) proceeded to analyze the roles of evaluation. He cited two main roles: formative, to assist in developing a program or other object, and summative, to assess the object's value once it has been developed. We note that it is not the nature of collected information that determines whether an evaluation is formative or summative but how it is used. If the information is used to guide development, the evaluation is formative. If it is used to sum up the value of something, the evaluation is summative. In these respects, the same data may be used for either formative or summative evaluation.

Evaluation in its formative application is an integral part of the development process. It provides continual feedback to assist in planning, developing, and delivering a program or service. In curriculum development, it addresses questions about content validity, the vocabulary level, usability, appropriateness of media, durability of materials, efficiency, staffing, and other matters. In classrooms, it may entail close and continuing assessment of teaching acts and each student's progress, with feedback used to strengthen both teaching and learning. In general, formative evaluation is done to help persons improve whatever they are developing, operating, or delivering.

In the summative role, evaluation

> may serve to enable administrators to decide whether the entire finished curriculum, refined by the use of the evaluation process in its first (formative) role, represents a sufficiently significant advance on the available alternatives to justify the expense of adoption by a school system. (Scriven, 1967, pp. 41–42)

Usually an external evaluator should perform a summative evaluation to enhance objectivity, and the findings should be made public. The summative evaluator searches for all effects of the object and examines them against the assessed needs of relevant consumers. He or she compares the costs and effects of the object to those of what Scriven has called "critical competitors," especially ones that might be less expensive and equally effective. In case the audience might be predisposed only to judge outcomes against the developer's goals, the summative evaluator provides judgments about the extent to which the goals validly reflect assessed needs. Overall, summative evaluation serves consumers by providing them with independent assessments that compare the merit, worth, and significance of competing programs or products.

Recently Scriven (2004a, 2004b) added a third major role of evaluations, labeled ascriptive evaluation. He identified ascriptive evaluations as not connected to a development process. An example that occurs to us is that a historian might conduct a retrospective evaluation of Henry Kaiser's use of competition between his California and Oregon factories, which miraculously decreased the production time of warships during World War II. This historian's evaluation would not be usable for improving Kaiser's employment of competition in the shipbuilding process (the formative role) or advising anyone about whether to purchase ships that Kaiser developed (the summative role). However, the historian's evaluation could yield an interesting analysis and judgment of Kaiser's use of competition in shipbuilding. In general, the same central logic of ascriptive evaluation applies to numerous other evaluative endeavors as well (Coryn & Scriven, 2008; Scriven & Coryn, 2008).

Although we have strained to illustrate Scriven's definition of ascriptive evaluation through this example, we think summative evaluation adequately covers this and other such examples. This is especially so if one espouses Scriven's original definitions of formative and summative evaluations as denoting uses of information rather than what information is collected and why, how, where, and when it is collected. We therefore think the term *ascriptive evaluation* might prove to be superfluous and disappear from the lexicon of evaluation concepts.

Amateur Versus Professional Evaluation

Scriven (1967) prefers, in the early stages of development, what he refers to as "amateur evaluation" (self-evaluation by persons with minimal evaluation expertise) over "professional evaluation." Developers, when they serve as their own evaluators, may be somewhat unsystematic and subjective; but they are also supportive, nonthreatening, dedicated to producing a success, and tolerant of vague objectives and exploratory development procedures. They are therefore unlikely to stifle creativity early on. Professional evaluators, if involved too early, may "dampen the creative fires of a productive group" (Scriven, 1967, p. 45); slow the development process by urging that objectives be clarified; or lose their objective perspective by becoming too closely aligned with the production effort, among other considerations. Professional evaluators are needed, however, to perform both formative and summative evaluations during the later stages of development.

Both formative and summative evaluations require high-level technical skills and objectivity seldom possessed by persons on the development staff who are not specially trained in the theory and methodology of evaluation. Scriven (1991, 1993) has recommended that a professional evaluator be included on the development staff to perform formative evaluation, and he has often advised that external professional evaluators be commissioned to conduct and report on summative evaluations.

Intrinsic and Payoff Evaluation

Scriven (1967, 1996) has distinguished between intrinsic evaluation and payoff evaluation. In intrinsic evaluations, evaluators appraise the qualities of a program, textbook, theory, or other object, regardless of its effects on users, by assessing such features as goals, structure, methodology, qualifications and attitudes of staff, facilities, public credibility, and past record. In payoff evaluations, evaluators are concerned not with the nature of the object, but rather with its effects on users. Such effects might pertain to test scores, job acquisition, job performance, or health status. Scriven acknowledged the importance of intrinsic evaluation, but emphasized that one must also determine and judge outcomes, because causal links between process and outcome variables are rarely, if ever, known for certain. He explained that both types can contribute to either formative or summative roles. He has often criticized accrediting boards for their preference for intrinsic criteria, such as the number of books in an institution's library, upkeep of facilities, and staff credentials and reputation, on the one hand, and their relative inattention to outcome variables, such as job success of graduates, on the other.

Goal-Free Evaluation

In yet another move against the widespread preoccupation with goals-based evaluation, Scriven (1973, 1974) introduced a counterproposal: goal-free evaluation. According to this approach, the evaluator purposely remains ignorant of a program's stated goals and searches for all effects of a program regardless of its developer's objectives. There are no side effects to examine,

because data about all effects, whatever a program's intent, are equally admissible. If a program is doing what it is supposed to do, then the evaluation should confirm this, but a goal-free evaluator also will be more likely to uncover unanticipated effects that a goals-based evaluator might miss because of his or her preoccupation with stated goals. Scriven (1991, 2007) has said that goal-free evaluation is reversible and complementary: one can start out using the goal-free approach to search for all effects and then shift to the goals-based approach to ensure that the evaluation helps determine whether goals were achieved, or both types of evaluation can be conducted simultaneously by different evaluators. Advantages of goal-free evaluation, according to Scriven (1973, 1974), are that it is less intrusive than goals-based evaluation; more adaptable to midstream goal shifts; better at finding side effects; less prone to social, perceptual, and cognitive biases; more professionally challenging; and more equitable in considering a wide range of values.

Goal-free evaluation is an innovative approach that is helpful in implementing the consumer-oriented approach to evaluation. In our evaluation practice, we have found that goal-free evaluation provides important supplementary information, expands the sources of evaluative information, is especially good in turning up unexpected findings, is a relatively low-cost procedure, and is welcomed and appreciated by clients (for example, Coryn, Schröter, Youker, & Bakerson, 2006; Schröter, Coryn, & Youker, 2006; Stufflebeam, Gullickson, & Wingate, 2002). Even so, some, including Patton (1997), among others (also see Shadish, Cook, & Leviton, 1991), have questioned the central premise of goal-free evaluation, and have sometimes referred to the approach as "goal-less" evaluation.

Needs Assessment

One challenge in using goal-free evaluation concerns how to judge an object's value based on the study's findings. If outcomes are identified without reference to what one is trying to accomplish, then how can one sort out desirable from undesirable consequences? Scriven's answer (1991, 1993, 2007) is that one must compare the observed outcomes to the assessed needs of the consumers. But if a need is a discrepancy between something real and something ideal, and if an ideal is a goal, then aren't needs assessments goals based, and, therefore, aren't goal-free evaluations also goals-based? Scriven says no. First, a developer's goals are not necessarily consistent with some set of ideals, such as those embedded in democracy. In any case, he maintains that the classic conception of a need as a discrepancy between something real and something ideal is wrong, because ideals are often unrealistic. Because the needs of consumers are a fundamental concept in his approach, he and others have extensively conceptualized and researched this concept (for example, Coryn, Gugiu, Davidson, & Schröter, 2008; Davidson, 2005; Stufflebeam, McCormick, Brinkerhoff, and Nelson, 1985).

For Scriven (Scriven & Roth, 1990), a need is anything essential for a satisfactory mode of existence, anything without which that mode of existence or level of performance would fall below a satisfactory level (see also Scriven, 1991). Some examples he has used are vitamin C and functional literacy. In the absence of these things, a person would be physically ill or socially and intellectually debilitated, respectively; hence, the person needs them. For Scriven, needs

assessment is a process for discovering facts about what things, if not provided or if withdrawn, would result in adverse consequences by any reasonable standards of good and bad. Given the results of such a needs assessment, an evaluator can determine the criteria and standards to be used in obtaining evidence for determining merit, worth, and significance, profiling critical competitors, and ranking or grading them (Scriven, 1991, 1993, 2007). Through this process, essentially the evaluator judges evaluation objects as good, bad, or indifferent (or, for example, ranks them first, second, or third) depending on how well each contributes to meeting identified needs. Scriven (1991) presented logical arguments for employing a needs-based approach to defining criteria and indicators and reaching evaluative conclusions, and he gave some leads on how to make this work in practice. In any given evaluation, however, much technical development will be required before needs assessment will offer a feasible means of defining evaluative criteria and standards and judging outcomes in a timely manner.

Scoring, Ranking, Grading, and Apportioning

Scriven (1991) has identified four main evaluative acts that are potentially relevant to all types of evaluation: scoring, ranking, grading, and apportioning (also see Coryn, 2007a). Scoring involves assigning numerical quantities to an evaluand or some aspect of an evaluand. These quantities each represent a sum of quality points, of which the points usually are assumed to be equal in value and additive. The range of possible scores usually is taken to represent lowest measure of merit to highest measure of merit. However, the value meaning of any single score is unclear without additional information.

In regard to ranking, two or more evaluands may be ranked based on their scores on a particular evaluative procedure. The relative ranks then indicate the merits of one evaluand relative to another, but not to particular levels of merit. Depending on whether the involved scale is ordinal, interval, or ratio, the distances between the scores of different evaluands may or may not be considered equal.

To obtain absolute judgments of merit, grades must be assigned to each possible score. For example, on a ten-point scale, grades might be assigned as follows: 0–2 = F; 3–4 = D; 5–6 = C; 7–9 = B; and 10 = A. Determining an appropriate range of scores for each potential grade requires examination of relevant information about the evaluand and similar evaluands; careful, systematic analysis of relevant evidence and logical arguments; and analysis of the nature and difficulty levels of items in the measurement device.

The fourth main method, apportionment, involves allocation of a finite set of resources to alternative evaluands; this typically involves scoring, ranking, and grading, which contribute to the final synthesis step (Scriven, 1994b). The case of a university that had to reconsider the doctoral programs it would support provides an example of apportionment evaluation.

A large, research-oriented university had insufficient funds to support all of its doctoral programs and needed to determine which programs should be discontinued and how available funds should be allocated to the remaining programs. The board of trustees stipulated that no tenured or tenure-track faculty members in discontinued programs would be dismissed; instead, they would be given opportunities to fill other open positions in the university

or to replace non-tenure-track staff. Given the evaluation's political nature, the university contracted with a highly credible external evaluator to coordinate and control the evaluation. The evaluation also included meaningful involvement and input from stakeholders, especially doctoral students and their professors. The evaluation used the procedures of scoring, ranking, and grading as predicates to the final synthesis and apportionment step.

The board contracted with an external panel to conduct a metaevaluation of this apportionment evaluation. The metaevaluation panel included a university president, a provost, a graduate dean, a college dean, a Nobel laureate professor, a doctoral student, and an evaluation expert. This team was tasked with evaluating the apportionment evaluation plan, the draft report, and the final report.

At the outset, the evaluator, the university's leaders, and a representative group of stakeholders compiled a set of criteria for evaluating and contrasting the programs fairly:

1. Need for the program's graduates as indicated by a record of more than a 70 percent rate of graduate employment in the subject discipline within one year of graduation.
2. Selection criteria and decision rules judged by external experts in the discipline to ensure acceptance of only high-quality students.
3. Rigorous application of the selection criteria, as evidenced by students' entry-level credentials, test scores, and previous grade point averages, plus judgments by external experts.
4. An acceptable number of students, defined as more than ten active students for each year of the program's existence (following the first year), up to and including the last three years.
5. High-quality students, with no more than 20 percent having a cumulative grade point average below a 3.5 during the previous year.
6. Acceptable graduation rates, defined as at least 80 percent of students graduating within four years of entering the program (with programs existing less than four years to be given a provisional pass on this criterion).
7. Qualified faculty, including at least three tenured faculty members who have been actively engaged with the program's students.
8. A duly approved curriculum that is judged positively by pertinent experts from outside the university.
9. High-quality courses, as judged by the program's students and graduates and by external experts.
10. Timely courses that students can take when they need them, to be judged based on interviews with students and examination of their records of courses compared with their approved courses of study.
11. Pertinent and rigorously conducted internships, as judged by the program's students and graduates and by external experts.

12. A noteworthy flow of grants and contracts providing students with meaningful practical and research experiences, as indicated by funded projects and students' positive judgments of their associated experiences.
13. High-quality dissertations as judged by external experts and as further indicated by spin-off publications.
14. An outstanding record of research by the program's faculty, as evidenced by grants and at least two noteworthy publications per year per faculty member.
15. A positive reputation of the program in its discipline, as judged by external experts in the discipline.
16. Cost-effectiveness of the program, judged as acceptable or not based on subtracting the program's annual amount of grant and contract funds from its annual cost to the university, dividing the difference by the number of program graduates, and comparing the per-graduate net cost with those of all of the university's other doctoral programs. The program's rank in this distribution of doctoral programs is to be considered when decisions are reached to retain or not retain the program.

Working from these criteria, the external evaluator and the university's provost divided the criteria into essential criteria (that is, items 1 through 7, 11, and 13) and important criteria (the remainder). The evaluator developed pertinent scoring rubrics. The evaluator and provost then defined decision rules for determining whether a program met or failed to meet each criterion, and the evaluator constructed a pertinent rating scale. The evaluator and provost also developed specifications and a schedule for each program to follow in preparing a portfolio of relevant evidence. Furthermore, they determined a plan and budget for commissioning panels of external experts to evaluate each program. The evaluator and provost reviewed this plan and budget with a group of university stakeholders appointed by the university's president. After making some modifications, the provost presented the plan to the president. She reviewed the plan with the external metaevaluation panel and, after securing some further clarifications and improvements from the evaluator and provost, approved the plan. Subsequently the evaluator and provost conducted orientation and training sessions for all groups that would participate in the evaluation.

In due course, the external evaluator obtained and reviewed the programs' portfolios and the assessments by external teams. He then scored and rated each program. He did so first in relation to the essential criteria. All programs that failed to meet one or more of these criteria were designated for termination—a minimum performance "bar" (Coryn, 2007a; Davidson, 2005; Scriven, 1991, 2007). In these determinations, the ranking was implicit in the partitioning of programs as acceptable and unacceptable. Next, the evaluator scored each remaining program on the remaining important criteria. Using preestablished decision rules, he subsequently converted the scores into ratings of excellent, good, or marginal. For each program, he prepared a profile of ratings on the important criteria and then computed an overall weighted average grade of excellent, good, or marginal. In addition, for each program, he appended an explanation of the rationale, information, and procedures used to arrive at the profile of ratings and the final grade.

Programs with an overall grade of excellent were assessed to determine their bottom-line funding needs. Basically these included costs associated with faculty and support positions, research associateships, materials and equipment, travel, research, and communication. The total of these funds was subtracted from the available funds. This process was repeated with the programs graded as good and subsequently with the programs graded as marginal.

Under this analysis, funds were not available to support continuation of any of the programs designated for termination pursuant to the first round of analysis keyed to the essential criteria or any of the programs graded as marginal in the subsequent analyses keyed to the important criteria. Following these assessments, the evaluator presented a draft of the results to the university's leaders, a representative group of stakeholders, and the external metaevaluation panel. After receiving criticisms of the draft report, the evaluator corrected factual errors, clarified areas of ambiguity, and submitted the final report to the university's provost and president. The provost then aired the report at a university-wide meeting, which stimulated a heated exchange. Subsequently the provost considered the issues raised, prepared his recommendations, and submitted them to the university's president. The president approved the report and submitted it to the board for action.

The board reviewed the report with the metaevaluation panel and concluded that it was sufficient and defensible. The board subsequently approved the elimination of all but one of the programs identified for discontinuation. It mandated the reform of one marginal program because it was in an area of high need and had many students. The board also decided to eliminate three programs that had been graded as good because they all had marginal ratings on criteria items 8, 12, and 16, and because the board judged that the funds being spent on these programs could be deployed more effectively elsewhere. The board directed the president to allocate the recovered funds to reform the one marginal program and strengthen two programs that had been rated excellent. The board also instructed the president to place the marginal program on probation; to direct that it dramatically improve on criteria items 8, 9, 12, and 14; and to put it under close scrutiny. Finally, the board directed the president to work out a process for phasing out the programs slated for elimination such that the university would meet commitments to existing students and tenured faculty.

This simplified example of an apportionment evaluation illustrates the differences among and general nature of the acts of scoring, ranking, grading, and apportioning and their functioning within a politically charged setting.

Checklists

Checklists, central to Scriven's methodological approach (1991, 2005b, 2007), contain relevant criteria for evaluating a particular object or conducting comparative evaluations of competing objects. Scriven constructs such checklists to guide the collection of relevant evidence and to grade or rank the one or more objects of an evaluation. He employs both generic and particularized checklists. The former include only the defined criteria for evaluating a class of objects. The particularized checklists include the defined criteria, rating levels, weights, and threshold standards of acceptability, all determined in consideration of an evaluation's context. His basis for constructing a generic checklist is an understanding of the nature and functional

properties of an object. In developing a particularized checklist, he requires close consideration of the operating context and the needs of both the client and the intended beneficiaries (Scriven 2005c; Stufflebeam, 2000b). (His definition of and rationale for evaluation checklists are available at www.wmich.edu/evalctr/checklists.)

Scriven (1994d) noted that constructing evaluation checklists is difficult due to the necessity of meeting such conditions as the following:

- Comprehensiveness in addressing all important criteria
- Nonoverlapping checkpoints to avoid double-weighting an area of overlap (that is, to avoid redundancy)
- A focus on direct measures of merit rather than statistical correlates of merit, although empirically validated indicators may be employed if direct measures are not feasible
- A consistent level of description for all checkpoints
- Amenability to operational definition and application to allow for a determination of whether and how well an entity meets a checkpoint

Key Evaluation Checklist

Scriven has continually refined his KEC, which was originally developed in the early 1970s for use by the Educational Testing Service in evaluating educational products produced by federally funded research and development centers in the United States (Coryn, 2006). The KEC is available in its most current form from www.wmich.edu/evalctr/checklists. This checklist, used to both plan, design, and evaluate evaluations,[1] is generic, but it can be adapted for use in particular evaluations. The checklist reflects Scriven's view that evaluation has multiple dimensions, must address pertinent values, should employ multiple perspectives, involves multiple levels of measurement, should use multiple methods, and usually should culminate in a bottom-line evaluative conclusion.

The checklist is divided into four major parts. Part A, "Preliminaries," contains an executive summary; a preface (examining the source and nature of the request or need for the evaluation); and methodology (for example, comparative or noncomparative; scoring, ranking, grading, or apportioning; or some combination of these). Part B, "Foundations," provides background and context, descriptions and definitions of the program and its components, consumers and other impactees of the program, program resources, and values (minimum standards and weights). Part C, "Subevaluations," offers evaluations of process, outcomes, costs, comparisons (alternatives), and generalizability. Part D, "Conclusions," comprises an evaluation of overall significance; possible recommendations and explanations; (possible) responsibility and justification (for example, allocating blame or praise); and metaevaluation.

These parts of the evaluation checklist need not be performed in any particular sequence, but all must be addressed or at least considered before the checklist has been implemented correctly. Also, an evaluator may cycle through the checklist several times during the evaluation of a program. Early cycles are formative evaluation; the last cycle is summative evaluation.

The rationale for the KEC is that evaluation is essentially a data reduction process, whereby large amounts of data are obtained and assessed and then synthesized into an overall judgment of value. In describing this data reduction process, Scriven (2007) has suggested that the early steps of an evaluation help characterize a program or product and the later checkpoints are intended to assess its validity.

The Final Synthesis

Scriven (1994b) presented extensive philosophical, theoretical, and methodological analysis of the final synthesis step in an evaluation. He noted that many evaluations that are impeccably conducted through the planning and data collection and analysis stages fail because an evaluator makes a leap from the data to a high-inference judgment, which often is only an idiosyncratic non sequitur. Such a judgment may reflect the evaluator's (or client's) biases more than the relevant background information and may stimulate mistrust of the evaluation rather than insightful action. Scriven (1994b) said:

> Sometimes . . . there is no way to avoid relying on judgment at this point. But, whether we call the last step clinical inference, intuition, professional judgment, connoisseurship, or impressionism, the solid body of the clinical vs. statistical research makes it clear that we rely on it at considerable peril. That research shows how a very simple rule, if it is empirically-based, can beat expert judgment—including the combined judgment of a panel of experts—in almost all cases. (p. 367)

What is needed, according to Scriven, are clear—but not simplistic—rules for deciding whether and how to reach justified conclusions. He has noted, however, that often these are not to be found.

One needs to pursue the final synthesis step under the assumption that it may be desirable and feasible to carry it out, and therefore one must proceed as far on this course as makes sense. The steps in the process are searching for an appropriate decision rule; deriving criteria admissible in probative judgments; deriving criteria of goodness inherent in the classical definition of the evaluation object; assessing the needs and preferences of the client and beneficiaries; obtaining evidence of each object's status on the criteria of merit, worth, and significance; weighting the criteria; profiling the results; deciding whether to try for a final synthesis; and, if warranted, combining the results to reach an overall conclusion.

Scriven (1994b, 2007) has urged evaluators to consider three factors in deciding whether to make a final synthesis:

- *Determine what the evaluation client needs.* Must the client have a final synthesis that compares critical competitors—all of which pass minimally acceptable standards on all significant criteria—on appropriately weighted criteria of merit? Or, for example, does the client need only a report card or profile on how each object rates on each criterion?

- *Consider the limitations of the available data.* If the available data on each object for each criterion are complete, and if the evaluator and client have been able to determine defensible weights and standards, the evaluator can both rank and grade the evaluation objects. Otherwise he or she should stop at the point of displaying each object's profile, taking into account the fact that the profiles may be incomplete.
- *Examine the configuration of available facts.* If the facts reveal a tie between one or more objects, the evaluator should report this and not try to pick a winner. Instead, he or she should advise the client to pick randomly or apply some additional, defensible criterion.

Scriven (1994b, 2007) has advised evaluators that in launching the final synthesis step, whenever possible they should get a valid rule in place by which different evaluators using the same data set would reach the same evaluative conclusion. If this is possible, the final synthesis step presumably is fairly straightforward. However, if such a rule is not to be found, which will often be the case, then the evaluator should, according to Scriven (1994b), try for heuristics and rubrics or, failing that, train and calibrate judges. If all of these steps fail or are seen in advance not to be feasible, then the evaluator should stop short of reporting a final, synthesized judgment of value and instead should deliver, for each evaluation object, a properly circumspect profile of performance levels on the significant criteria that guided the evaluation.

Scriven (1994b) recommended several general procedures for reaching the final synthesis. One is "probative inference," which is "inference that makes a prima facie case for a conclusion: the kind of inference . . . which is highly contextual" (Scriven, 1994b, p. 371). Such inference involves deriving values from facts. According to Scriven (1994b), evaluative claims are facts if they are extremely reliable judgments made by experienced judges against simple, valid standards—judgments that are scrutinized, consciously or unconsciously, by a trained evaluator for errors of fact, standards, judgment, or inference.

Another procedure that can be used as part of the final synthesis step involves using definitional claims in determining criteria and ultimately reaching an evaluative conclusion (Scriven, 1994b). For example, maintaining discipline in the classroom is generally regarded as an essential component of the definition of good teaching, which makes effective classroom management an essential criterion for evaluating teaching. Examples Scriven gave are "Good watches . . . keep good time" and "Good judges . . . do not take bribes" (p. 372). Because these definitional claims can be supported by factual and analytical evidence about use and the way watches and judges work, respectively, they are significant, nonarbitrary criteria for evaluating watches and judges. Under this approach, the evaluator defines and defends the criteria of merit by determining the factual claims inherent in a concept of interest and identifying known, relevant facts about objects covered by this concept. These functional analysis steps provide a foundation for comparing evaluation objects on the selected criteria and for judging each object in absolute terms against standards of what should be expected of that object.

Another frequent component in the synthesis process involves assessments of beneficiaries' needs and wants:

> Assuming that the client's problem calls for a ranking, then you must turn to the needs and—to a lesser extent—the preferences of the targeted recipients to provide the . . . relevant criteria and their weights . . . Typically, the procedure would require you to try for a comprehensive list of criteria and tentative weights based on the concept, the experience of the service providers, and the literature, and take that to the consumers for additions. (Scriven, 1994b, p. 377)

As a last point about the final synthesis step, Scriven (1994b) warned against the fallacy of the numerical weight and sum (NWS) approach to reaching an evaluative conclusion (also see Davidson, 2005). This relatively common approach involves computing an overall score on an evaluation object by summing across all criteria the products of each criterion's weight times the object's score on the criterion. This procedure could erroneously give a passing grade to an object that failed or did poorly on the most important criteria but scored high on less important or even trivial criteria. To replace this faulty synthesis procedure, Scriven (1994b) offered the qualitative weight and sum (QWS) approach. Using this approach, the evaluator begins by rating the evaluative criteria on their significance as essential, very important, important, just significant, or not significant. The not significant criteria are dropped. Then the evaluator immediately drops from further consideration any object that fails to pass the essential criteria. According to Scriven, the essential criteria then become a moot point because the remaining evaluation objects meet them. (We see this conclusion as not always the case, as explained beginning in the next paragraph.) Subsequently the evaluator develops for each remaining object three scores, representing the number of criteria passed in each remaining significance group (very important, important, and just significant). Keeping the three scores separate means that "no number of points scored in the currency of lower-weighted criteria can overpower points picked up on a higher-weighted criterion" (Scriven, 1994b, p. 376).

We see Scriven's QWS procedure as creative and useful but too restrictive. Sometimes a simple sum of criteria met in a category will be sufficient to judge an object's merit in the category, and sometimes not. We would not necessarily drop consideration of essential criteria following the discarding of objects that failed on these (although we recognize that the remaining objects would, according to Scriven's procedure, all have a score equal to the category's number of criteria). Nor would we categorically reject an NWS approach to arrive at scores of merit. If a rating scale is applied to each criterion in a category, some objects will surely score higher than others on given criteria and on the total of criteria in the category. We think such variations often will be of interest to decision makers. Although Scriven (1991, 1994b, 2007) has not provided for rating each object's relative level of satisfaction of each criterion, we think this possibility should be preserved, although not required. If each object is rated on, say, a three-point scale for each criterion (1 = pass, 2 = pass with distinction, and 3 = exemplary), then a total category score could be obtained for each object by summing the ratings and dividing by the number of criteria in the category. Keeping each set of category scores separate

would preserve Scriven's objective of ensuring that points on lower-weighted criteria would not overpower points earned on higher-weighted criteria. Rating each object on criteria within a category would also reveal each object's relative rated merit within the category.

A weight and sum approach could be employed next to obtain an overall score for each object. For example, the evaluator could compute an object's overall score by first weighting each category of criteria (for example, 4 for essential criteria, 3 for very important criteria, 2 for important criteria, and 1 for just significant criteria). Then, for each object and category of criteria, the evaluator would multiply the category's weight by the object's normalized category score. Next, the evaluator would sum each object's four weighted category scores and divide by four. The evaluator could subsequently rank all objects still under review on their derived total scores. A report showing ranks of objects within and across categories of criteria would prove useful for drawing conclusions and making decisions about the different objects. We believe that decision makers often will want to see how objects that made the initial cut ranked within the categories of essential, very important, important, and just significant criteria, and overall. We suggest the procedure we have outlined as a perhaps useful extension of Scriven's QWS approach. This extension might be labeled the qualitative and numerical weight and sum approach (QNWS).

Metaevaluation

The final item in the KEC calls for the evaluation of evaluation. Scriven introduced this concept in 1969, when he published an article responding to questions about how to evaluate evaluation instruments. He cited this as one of many concerns in metaevaluation and emphasized that evaluators have a professional obligation to ensure that their proposed or completed evaluations are subjected to competent evaluation. His rationale was that evaluation is a particularly self-referent subject because it "applies to the process and products of all serious human endeavor and hence to evaluation" (p. 36). He noted that metaevaluation can be formative, in assisting the evaluator to design and conduct a sound evaluation, or summative, in giving a client independent evidence about the technical competence of the primary evaluator and the soundness of his or her reports. Scriven's methodological suggestions (2007) for conducting metaevaluations include the use of his KEC to assess an evaluation as a product; the use of some other checklists (for example, Scriven, 2011b); and the use of professional evaluation standards.

Evaluation Ideologies

Scriven has been one of the most thoughtful and vocal critics of prevailing views of evaluation (1983, 1993). Consistent with this critical stance, he has emphasized that evaluation is a particularly self-referent subject, which adheres to his advocacy of metaevaluation. He has classified these prevailing views into four groups and critiqued each extensively in the hope of convincing evaluators to recognize and shed certain biases, which he claims have debilitated evaluation work. And he has used his analysis of strengths and weaknesses of each approach

to strengthen his rationale for the KEC, describing this checklist as one that encompasses the best features of all other serious proposals about how to do evaluation and avoids the flaws he identified in the other proposals. Further insight into Scriven's philosophy of evaluation in general and the KEC in particular can be gained by carefully considering his analysis of alternative ideologies (Scriven, 1983). Therefore, we next capture his most salient points in regard to each of four ideologies: the separatist, positivist, managerial, and relativistic ideologies.

Separatist Ideology

Scriven sees the separatist ideology as rooted in the denial or rejection of the proposition that evaluation is a self-referent activity. This ideology is best reflected in evaluation proposals that require the appointment of evaluators who are totally independent of what is to be evaluated. Establishing and maintaining an evaluator's independence from an evaluand is often seen as essential for ensuring that evaluation reports are unbiased (Scriven, 1975, 1983, 2011e). In addition, evaluators who practice this ideology, according to Scriven (1983), often fail to recognize or address the need to have their own work evaluated. Quite possibly many of them see such metaevaluation as a concern for somebody else, because evaluators, according to their separatist view, could not be objective in evaluating their own work. Hence, Scriven (1969b) pointed to the paradox of an evaluator who earns a living by evaluating the work of others but fails to see, or may even resist, evaluation of his or her own services.

Underlying this kind of professional parasitism is, according to Scriven (1983), a basic human flaw: "valuephobia, a pervasive fear of being evaluated" (p. 230; also see Donaldson, Gooler, & Scriven, 2002). It is manifest when evaluators who are in close contact with the person whose work they are evaluating become co-opted, lose their critical perspective, and praise what they might have criticized had they maintained greater distance from the evaluand. Valuephobia may also be present when evaluators resist, or at least avoid, having their evaluations evaluated.

In opposing the separatist position, Scriven (1983) argued that professionals, including professional evaluators, need to acknowledge and deal straightforwardly with the self-referent nature of evaluation. The hallmark of professionalism is subjecting one's work to evaluation. The fact that all evaluations are prone to bias should not deter one from evaluating one's own work or commissioning someone else to do so. Instead, one should respond by conducting the evaluation in as unbiased a manner as possible and subjecting the evaluation to scrutiny against recognized standards of sound evaluation. Further, in program evaluations, evaluators should look realistically at program staff as well as other aspects of a program, because success and failure invariably are inseparable from the work of staff and there will be little prospect for improvement through evaluation if guidance for improving the performance of staff is not provided.

It is of interest in regard to the preceding discussion that the Joint Committee on Standards for Educational Evaluation introduced a new, fifth category of standards—labeled evaluation accountability—in its latest edition of *The Program Evaluation Standards* (2011).

This category includes standards for evaluation documentation, internal metaevaluation, and external metaevaluation. Basically, these new standards are consistent with Scriven's call (1983) for self-referent evaluation by professionals. We see this as especially so concerning the first two of these standards, in which evaluators are held responsible for documenting and attesting to the merit of their evaluations. However, contrary to the Joint Committee's recommendations and what Scriven might endorse, we think it is inappropriate, as regards the external metaevaluation standard, for the evaluator to select, engage, fund, and control the quality of the external metaevaluator. Such practice is prone to an evaluator's bias in seeking out a "friendly critic"—one who can be expected to issue a positive report, whether merited or not. We think the evaluator should advise the client to assume responsibility for securing an independent, external metaevaluation and that the evaluator should fully cooperate with but not in any way control the external metaevaluation. Nevertheless, we endorse Scriven's main point (1983, 1991, 2007) that evaluations themselves should be validly evaluated.

Positivist Ideology

Scriven (1983, 1991, 1993) has identified a second ideology, that of logical positivism, as another overreaction to valuephobia. He has argued that positivists, in their attempts to remove bias from scientific works, overreact to the point of trying to render value-free both twentieth-century science in general and evaluation in particular. Whereas the separatists reject the self-referent nature of science or evaluation, the positivists reject the evaluative nature of science. Scriven pointed to a number of contradictory cases—for example, educational psychologists who assert that no evaluative judgments can be made with objectivity yet easily produce evaluative judgments about the performance of their students. Scriven's response (1983, 1991, 1993) to the flaws of positivism has been to give central importance to the practice of assigning value meanings to the findings obtained in evaluation studies. Scriven's recent call (2011c) for three revolutions in the evaluation field entails a strict rejection of the value-free ideology.

Managerial Ideology

For Scriven (1983), a "well-managed evaluation" often means much more than one that is guided by a competent evaluation administrator. It can instead involve "a very self-serving indulgence in valuephobia" (p. 238) by both program managers and evaluators. A program manager may impose rigid controls over the evaluation he or she commissions so that there will be no surprises. The program manager may want only his or her program evaluated, not the personnel who operate it and especially not its administrator. And the program manager might insist that the evaluation be limited to determining whether his or her stated goals for the program have been achieved and that the evaluator be restricted from judging the program manager's work based on somebody else's wishes for the program.

From the program manager's perspective, this managerial ideology clearly includes a bias toward producing favorable reports. According to Scriven (1983, 1993), many evaluators are willing to fulfill the manager's wishes for favorable, predictable reports because of a parallel

set of self-serving reasons. They want future contracts or to retain their position as evaluator in an institution, and giving a favorable report, or at least one that does not make their client and sponsor nervous, is most likely to lead to obtaining future work. They are often willing to partial out any concern for personnel evaluation because this helps make the evaluation more feasible as well as independent of the different and often conflicting value positions that different persons involved in the program might hold. The manager's request for limiting the assessment to what had been intended is especially congenial because it means the evaluator adhering to this ideology will probably avoid not only having to assess the implementation of the program and especially the performance of staff members in it but also having to deal with values, because they are presumed to be given in the program manager's goals.

In the managerial ideology, then, we can see the possibility of a confluence of the separatist, positivist, and managerial ideologies—all with bad effect. By avoiding evaluation of the manager and staff (consistent with the separatist ideology), keeping the evaluation as a technical service devoid of value determinations (the positivist approach), and helping the manager get the good report he or she needs on the accomplishment of his or her goals (the managerial ideology), the evaluator has effectively caused evaluation to be a disservice rather than a contribution to society.

With the bent just outlined, the study, according to Scriven (1983), would exclude many vital aspects of a sound evaluation. It would deter the client from rather than assist him or her in examining goals and services critically. By concentrating on a developer's goals, the evaluator would fail to ensure that a program has value for addressing consumers' needs. The study would probably be myopic and not consider whether the program is a "best buy," when it could serve the client better by exposing and comparing alternatives. And it would be likely to skirt issues concerned with ethics and prudent use of scarce resources. For Scriven (1983, 1991, 1993), the widely seen adherence to the managerial ideology and the ideology's connections to other bad evaluation practices are a travesty for society and for the evaluation profession. He has used his critical analysis of this stance as a platform from which to advocate a series of reforms, which are seen in his KEC:

- Performing needs assessments as a basis for judging whether a program has produced beneficial outcomes
- Evaluating "goal-free" so as not to become preoccupied with the developer's goals and thereby miss finding important but unanticipated outcomes, good and bad
- Comparing what is being evaluated to viable alternatives
- Examining services for their cost-effectiveness
- Combining personnel and program evaluation

Relativistic Ideology

Another ideology that Scriven (1983, 1991, 1993) sees as flawed and debilitating in its influence on evaluation work is the relativistic ideology. Scriven considers it to be an overreaction to

problems associated with the positivist ideology. Whereas the positivists often have put forth the view that there is an objective reality that can be known by anyone who can and will use unbiased assessment procedures, the relativists have charged that this construction is overly simplistic and can only lead to narrow assessments that give exclusive and undue prominence to the perspective of some group in power under the mistaken view that its perspective and assessments are objective. In response to the hazards of positivism, the relativists assert that all is relative, that there is no objective truth. Therefore, they call for multiple perspectives, criteria, measures, and answers.

According to Scriven (1983, 1991, 1993), this movement in the evaluation field sometimes denies the possibility of objective determinations of merit or even objectively correct descriptions of programs. Although he also rejects the existence of a single correct description, he counsels us not to abandon the idea that there is an objective reality. It may be a complex reality beyond our existing capabilities to comprehend and describe thoroughly, but we only delude ourselves if we pretend it does not exist. He counsels instead that we may need to relativize our descriptions for different audiences. But he cautions us not to accept all conflicting descriptions as correct as, we think, some of the more pedantic relativists seem prone to do. Instead, we are advised, as evaluators, to seek out the "best," the "better," the "ideal."

Avenues to Causal Inference

The past forty years have seen substantial controversy in the evaluation field related to the concept of causal conclusions in evaluations (Cook, Scriven, Coryn, & Evergreen, 2010). Campbell and Stanley (1963) argued that researchers and evaluators should assess the extent to which a project has caused observed outcomes and that the best way to obtain valid findings about cause and effect is through rigorous application of randomized controlled experiments. Since then, the U.S. government has repeatedly designated randomized controlled experimental design as the gold standard for evaluations and often has mandated this approach for use in federally funded evaluations (Donaldson, Christie, & Mark, 2009). Such requirements have held hostage much of the federal funding available for evaluations, with government officials releasing these funds only to evaluators who agree to conduct randomized controlled experiments. This practice directly opposes the advice of Fisher (1951), the father of experimental design. He expressly directed inquirers not to equate his experimental methods with science. Consistent with Fisher's position, many evaluation and research methodologists have sharply criticized the U.S. government's requirement that such experiments be used, arguing that nonexperimental methods often are superior to randomized controlled experiments in taking account of the real-world circumstances of programs, detecting causal relationships, and creating deep understanding of programs. Nevertheless, a number of leading researchers have continued to convince federal government officials that only sound controlled experiments can provide defensible conclusions about causes and effects in education, health, and human service programs.

Scriven (2005b) has strongly criticized the near monopoly that experimental design holds over federally funded evaluation, positing that the gold standard for evaluations should not

be experimentally determined conclusions but conclusions beyond reasonable doubt. Arguing that experiments often are inadequate and not even the best way to address questions of cause and effect, he stated, "We must agree that cause is an epistemological primitive, well understood by humans but not entirely reducible to other logical notions such as necessity and sufficiency" (p. 44). Referring to apparent causal linkages between the execution of a program and observed outcomes, Scriven (2005b) said that equal or better outcomes might have been produced by another program not under review or that, outside of the study's controlled conditions, the experimental program might not produce the same outcomes when applied in a naturalistic setting. The latter point gives credence to those who argue for in-depth case studies of programs in their natural settings. Scriven (2005b) concluded that the claim that randomized controlled experiments are consistently superior in identifying necessary and sufficient conditions that caused observed outcomes is a fallacy whose exposure should clearly make room for a range of methodological approaches to studies of cause and effect.

In advancing the range of methodological approaches that can meet his gold standard of conclusions beyond reasonable doubt, Scriven (2005a) looked to both quantitative and qualitative designs. He credited the following quantitative designs, when suitably applied, as meeting his gold standard: double-blind studies, single-blind studies, randomized controlled experiments, regression discontinuity studies, strong interrupted time-series studies, and identical-twin studies. Looking beyond these quantitative designs, Scriven also identified designs in the broader sphere of scientific inquiry: laboratory investigations used in forensic studies of the cause of a death; engineering studies of the cause of a structural failure in, for example, a bridge, a building, or a jet airliner; astrophysicists' conclusions drawn from systematic applications of astronomy methods and anthropological methods such as those applied by Darwin; and historical analysis. (We would add epidemiological studies and gnotobiotic studies of the effects of specified agents on germ-free animals.) After identifying this range of designs that can meet this gold standard, Scriven (2005a) cited randomized controlled experiments as only one of at least eight scientifically acceptable approaches to studying and reaching conclusions on causes and effects. In a recent dialogue with Cook, Scriven discussed what he perceived as the three claims associated with the superiority assigned to randomized experiments for causal investigations (Cook et al., 2010, p. 116):

A. The claim that only the RCT [randomized controlled trial] design excludes all alternative causes: This claim fails because it is agreed that it is only true if RCTs are double-blind, and in human investigations, that condition is rarely met. Note A1: In fact, it is arguably only true for triple-blind studies that cast doubt on most drug studies, and most human behavior studies.

B. The claim that only the RCT design supports the counterfactual analysis of causation: This claim is irrelevant because the counterfactual analysis is invalid; it is refuted by all cases of overdetermination, a common situation in medicine and human affairs.

C. The claim that no other design establishes causal claims with comparable certainty: This claim is false, since it is refuted by several other approaches, for example:

1. Critically filtered observational claims, for example, damage caused by an explosion seen by witnesses; color change in flask from adding litmus solution; making a noise by clapping
2. Highly scientific claims in forensic sciences, for example, autopsy reports
3. Highly scientific claims in case study research, for example, modified Success Case Method (MSCM; Coryn, Schröter, & Hanssen, 2009)
4. Many good quasi-experimental designs, for example, interrupted time series, especially with random time of treatment application and duration
5. Many theory-based causal claims in science, history, and law, for example, that the collision of tectonic plates caused the Sierra Nevada, that a meteorite caused Meteor Crater, AZ and Tycho on the Moon; smoking causes cancer, etc
6. Many (other) applications of the general elimination method (GEM)

Subsequently, Scriven (2005b) presented his final synthesis in regard to an appropriate general methodology for identifying and crediting a program's effects. Seeing observation as the most important and reliable source of causal claims, he stated, "Basic kinds of causal data, vast quantities of highly reliable and checkable causal data, come from observation, not from elaborate experiments" (p. 46). He said this conclusion liberates such field sciences as biology and anthropology from second-class citizenship in reaching and defending claims based on observed causal connections. Moreover, he said, rigorous applications of case study methodology can produce defensible causal claims. He stressed, however, that all studies of cause and effect should be subjected to strong methods of verification. Among others, these methods are independent confirmation, valid triangulation, detailed documentation, systematic elimination of alternative possible causes, and systematic qualitative analysis of causal chains and patterns. In the end, Scriven (2005a, 2005b) has acknowledged that development of valid causal claims is complex and difficult. Accordingly, he has called for an approach to studying causal connections that is grounded in systematic observation, involves the application of multiple methods for determining cause, and meets demanding standards of scientific explanation.

Product Evaluation

In 1994 Scriven (1994d) wrote an article in which he presented what he considered to be the state of the art in product evaluation. (Note that Scriven's use of products refers to produced items, such as refrigerators, textbooks, computer programs, and films, whereas Stufflebeam, in his Chapter 13 discussion of product evaluation, referred to products as the outcomes of a program.) Essentially, this article either encapsulates or alludes to some of the important concepts that Scriven had developed over the years and that have been reported in this chapter: needs assessment; goal-free evaluation; standards of various kinds (particularly those that refer to ethical, legal, or political considerations); metaevaluation; and the managerial ideology (particularly in respect to aspects of the KEC). We have extracted and summarized key aspects of this article.

The Place and Importance of Product Evaluation

Scriven began his article (1994d) in this way:

> Product evaluation is important for several reasons. The obvious one, which makes it sometimes a life-saving matter for the consumer, arises because our lives, and the quality of those lives, depend on the evaluation of products by external agencies—for example, on the evaluation of drugs and of automobile safety systems. The second, a (metaphorically) life-saving matter for inventors, manufacturers, and service providers, is the role of product evaluation in the improvement of products and services, a role which has, for example, driven the computer field to an unmatched rate of improvement, although the quality of its product evaluation leaves much room for further improvement. The third reason for its importance is its involvement within other applied fields of evaluation, particularly within program evaluation.
>
> The extent of this involvement is only now beginning to be appreciated, just as the extent of the involvement of personnel evaluation within program evaluation is only now emerging. An important example comes from the evaluation of programs using educational technology to improve instruction . . .
>
> It is rare that these programs can be evaluated without serious evaluation of the technology itself; yet it is also rare that the evaluation of the technology manages to avoid falling into the trap of using expert reviewers with shared bias or using independents who make invalid commentary from ignorance. The fourth consideration is that product evaluation has long served as an exemplar for other applied fields in evaluation, an effect which has been considerable and could still be extended with profit . . . The goal-free approach to program evaluation is an example. (p. 45)

The article offers general remarks about the field, followed by analysis of strengths and weaknesses of product evaluation by leading practitioners in consumer products, the automobile industry, and computers. Particular emphasis is given to Consumers Union (to be discussed shortly).

Despite the fact that product evaluation (and, to a lesser extent, personnel evaluation) has long been practiced—perhaps longer than any other type of evaluation (Scriven, 1991)—its methodology has not received the same attention as program evaluation. In the past fifty years, however, product evaluation has become considerably more developed and certainly more public. Scriven (1994d) commented that the increase in extent is exemplified by the development of technology assessment and an emerging literature on evaluating medical tests. The most obvious indication of the growing sophistication of product evaluation is the extensive array of magazines and newsletters devoting space to the results of product testing. Moreover, as Scriven (1991, 1994d) pointed out, the burgeoning of the field is also exemplified by growing emphasis on comparative rather than stand-alone tests. The consumer benefits by these activities; both utility and validity of choice improve. The overall outcome is an extremely useful, although often complex, set of resources—"if you know how to get to them and how to use them" (Scriven, 1994d, p. 46).

These improvements in product evaluation for the consumer do not, however, necessarily reflect better product evaluation by manufacturers and vendors. Poor product evaluations have often resulted in poor-quality products and services. Scriven (1994d) gave a prime example in which the U.S. automobile industry steadfastly refused to improve the quality of products, despite adverse criticism by external product assessors. He pointed out that as a result, there was a steady decline until the industry itself took stock of its position in relation to the price, reliability, and performance of foreign vehicles.

Basic Methodology

Answering the question of how product evaluation should be done, Scriven (1994d) advised that the same general formula as that used for all other evaluations should be followed. This entails identifying and validating criteria of merit, determining performance on those criteria of judgment, and integrating the two (that is, the criteria and performances) on the basis of some valid principle. Scriven maintained that "all the skill lies in the application of that formula" (p. 47). Specific details of a product evaluation will vary according to the type of product being tested and exact knowledge of specific consumer needs. Scriven (1991) stated that as a methodology, evaluation is a transdiscipline,[2] and developing this methodology poses a dilemma faced by the pursuit of logic (another transdiscipline) over two millennia. One can attempt to provide a general model, but this often turns out to be too difficult to apply reliably to other cases. Or one can focus on weaknesses, or "traps," known in logic as the "fallacies approach." Scriven sees the fallacies approach as possibly the most convincing to use in establishing a methodology for product evaluation.

How well is product evaluation done? Scriven (1994d) noted in his article that most of the virtues had already been given, so he focused on the shortcomings (in line with the fallacies approach). The shortcomings, which are serious, are most often seen in the faulty use of the formula for good evaluation, particularly in the incorrect application of widely validated principles to specific cases. In that respect, the problems are like those in most applied fields in evaluation; in ethics, for example, it is usually not the Ten Commandments (or their equivalent) that are in dispute, but how to apply them to specific cases.

Consumers Union

Scriven (1994d) noted the importance of *Consumer Reports,* the official organ of Consumers Union (CU), and acknowledged its status as a standard-bearer, albeit de facto, for product evaluation. However, he also presented a litany of shortcomings in the magazine's approach to product evaluation. He observed that it is no surprise that in the course of becoming extremely powerful, wealthy, and the principal institution of its kind, CU has also become almost immune to serious criticism. As a result, he said, its earned respect as a "near flawless paradigm of the state of the art" (p. 48) has slipped. For example, although CU has steadfastly excluded advertisements of the objects evaluated in its magazine, it began advertising its own products in its own magazine. Scriven saw this as "a change which surely tends to shift its value system near to that of the advertiser rather than the consumer or evaluator" (p. 48). Another cited

mistake was frequent use of the fallacious numerical weight and sum model for the synthesis of subevaluations. Scriven (1994d) stressed that this is a serious error in any evaluation, including product evaluation, when failure in a single component could override success in all the others. Looking at the other side of such criticisms, Scriven (1994d) noted that *Consumer Reports* generally remains an invaluable resource for consumers, because most of its findings are based on good work.

Professionalization of Evaluation

The description of Scriven's article (1994d) on product evaluation gives some indication of the extent to which he has been a main force for developing evaluation theory and methodology and professionalizing the evaluation field. His interests and contributions have covered virtually all aspects of evaluation, and his development of these aspects has been influential. As its first president, he helped to establish and develop the Evaluation Network, a professional organization for evaluators from education, health, government, and social programs and one of the two organizations that merged to become AEA. He developed the Evaluation Network newsletter into the highly respected publication, *Evaluation News,* which evolved into the *American Journal of Evaluation,* and recently he led development of the online *Journal of MultiDisciplinary Evaluation*. He was the 1999 president of AEA. His *Evaluation Thesaurus,* the fourth edition of which was published in 1991, is an important compilation of theoretical and philosophical ideas about evaluation. He has supported the Joint Committee in the production of professional standards for evaluations (see Chapter 3), but he criticized the committee for generating standards only for evaluations of programs, personnel, and students. He noted that given evaluation's ubiquitous nature, such standards are needed for evaluations in every discipline and area of service. Accordingly, we wonder if he would judge that the American National Standards Institute (ANSI) is responding sufficiently to this need for evaluation standards, given that ANSI has approved more than ten thousand such national standards, including those of the Joint Committee. As noted previously, Scriven directed the Western Michigan University IDPE program, thereby carrying out his position that evaluation has wide applications across disciplines and service areas. Scriven is the well-deserving recipient of many awards and prizes.

Scriven's Look to Evaluation's Future

Scriven was duly honored (in August 2011) at a Claremont Graduate University symposium to celebrate his outstanding career and contributions to the field of evaluation. For that occasion he wrote a provocative paper (Scriven, 2011a) titled *Conceptual Revolutions in Evaluation: Past, Present, and Future*. The piece builds on rather than departs from his past legacy of contributions to the theory, methods, and professionalization of evaluation. In this paper, as he takes stock of evaluation's past, present, and future, he describes and discusses three needed revolutions in the transdiscipline of evaluation.

The Evaluation Profession's First Revolution

In introducing the first revolution (R1), Scriven (2011a) stated that it is not yet complete and involves the need for all evaluators (1) to reject the view that evaluations should be value-free and (2) instead to embrace, not only in their rhetoric but in their actions as well, the position that evaluation practice should and can be values based and scientifically legitimate. He posited that the bottom-line consequence of R1 is that a competent evaluator—employing valid criteria and rigorous methods—can show that an evaluand is truly excellent or truly worthless as a matter of scientific fact. Scriven noted that although many evaluators espouse this concept of evaluation and manifest it in their practice, too many others say the right words but nevertheless persist in pursuing evaluation projects according to a value-free doctrine of science. This, he says, must stop, because it diverts evaluations from addressing the basic value questions of good, better, and best that are essential for assessing and judging efforts to address great societal needs.

The Evaluation Profession's Approaching Second Revolution

Referring to what he sees as the second revolution in evaluation (R2), Scriven (2011a) posited that evaluators are shifting their concept of evaluation from that of merely a respectable discipline to that of the alpha or first-order discipline within every other discipline. He sees evaluation not only as a transdiscipline that itself has many subdisciplines (for example, evaluation of products, programs, personnel, and organizations) but also as a discipline that is essential to assessing, judging, and validating work in every other discipline (such as sociology, psychology, anthropology, agriculture, economics, education, political science, medicine, and engineering). In illustrating the need for the social sciences and other disciplines to accept and deploy professional evaluation as the alpha subdiscipline in their field, Scriven (2011a) observed that peer reviews are essential to ensuring sound practices in any discipline or profession. Although he acknowledged that peer reviews are widely applied, he also charged that they are notoriously unreliable and likely to be seriously counterproductive, even injurious. Here, he sees a clear indication that each discipline should include and value the contributions of a competent evaluator on every peer review committee. In general, he noted that disciplines should beneficially engage teams that include members with the pertinent disciplinary expertise and one or two other members with highly honed evaluation expertise.

The Evaluation Profession's Eventual Third Revolution

Scriven (2011a) envisions a third revolution (R3) that will build on the first two revolutions but reverse the status of evaluation from one of just the first-order or alpha subdiscipline within other disciplines to one of a "paradigm discipline" for (at least) the social sciences. This is a sweeping projection that is seen to span about a century. Envisioning a time when this transformation is complete, Scriven sees each affected discipline as being essentially bicameral in structure, one division grounded in evaluation concepts and methods, the other grounded in the content of the field. This is reminiscent of Cronbach's classic article (1975) titled "Beyond the Two Disciplines of Scientific Psychology"—one relating to the content of

psychology, the other to the concepts and methods of evaluation and research. In characterizing R3, Scriven (2011a) noted that any discipline's credentials have two dimensions: disciplinary quality (the domain of evaluation) and content or territorial validity (the domain of the field's substance). The long-term recommendation here is that once R1 and R2 have firmly established sound evaluation methodology—across all evaluation subdisciplines, the social sciences, and possibly other disciplines—a given discipline should adopt rigorous evaluation methods and employ qualified staff members to constitute its second main, evaluation division. Accordingly, the involved disciplines are projected to have nonevaluative (content related) and evaluative scopes of work. The work of the evaluation division would be enormous and critically important, including evaluations of programs, personnel, products, training and education, research, and so on, plus continuing education of the content division's members in the proper concepts, methods, and uses of systematic evaluation. Scriven (2011a) has stressed that the reorganization suggested in R3 is not trivial, because leaders in most social sciences are uneducated in regard to the proper concepts, methods, and uses of sound evaluation, as it applies to all aspects of their field. They need to be taught to see and overcome errors of the past, including (1) working from unjustified assumptions; (2) failing to address the most important values-oriented questions (such as those concerning good, better, and best, and also those concerning goals, needs, cultural fit, ethicality, side effects, costs, and sustainability); and (3) counterproductively following a value-free doctrine of inquiry. In concluding his paper, Scriven (2011a, 2013) claimed that R3 is going to take a great deal of work as well as a great change in our conceptualization of the nature and role of both evaluation and (at least) the social sciences. Evaluators and members of disciplines, according to Scriven, need to recognize that any social science has limited value without including and making systematic use of evaluation as a core component. Clearly, with his paper Scriven (2011a) offered rich food for thought for those who wish to accept the challenge of his 3^F and pursue projects to accelerate the further development of the evaluation transdiscipline.

At this writing, we have just received the book based on the Claremont Graduate University symposium that honored Scriven. Edited by Donaldson (2013), the book includes, in addition to the chapters by Donaldson and Scriven, chapters by some of evaluation's most prolific authors: Rodney Hopson, Michael Patton, Ernest House, Daniel Stufflebeam, Christina Christie, Robert Stake, Jennifer Greene, Karen Kirkhart, and Melvin Mark. The ten chapters provide a rich view of the iconoclastic contributions of one of the evaluation field's most productive trailblazers.

Summary

This chapter has provided an overview of Scriven's many conceptual, methodological, and profession-building contributions to evaluation. Typically, his philosophically based recommendations have been countermeasures to what he sees as wrong and dysfunctional in traditional views of evaluation, especially a focus on developers' goals. He has sharply criticized both classical and more recent conceptualizations of evaluation. He has grounded his consumer-oriented view of evaluation in basic philosophical propositions of objectivism and pragmatism, and he has evolved concepts and methods to help articulate and apply

his approach—especially needs assessment as a key basis for judging outcomes. He also has been one of the foremost leaders in the effort to professionalize evaluation work, develop the evaluation discipline, and extend its application and contributions into all other disciplines.

By following Scriven's philosophy, an evaluator seeks to find those approaches that best address the assessed needs of consumers. Scriven has developed many key evaluation concepts, including apportionment evaluation, ascriptive evaluation, metaevaluation, and goal-free evaluation. His main approach encompasses formative evaluation keyed to helping develop sound programs and products and summative evaluation that assesses the merit, worth, and significance of developed products and services. He developed the Key Evaluation Checklist for applying his approach, as was illustrated earlier. He has argued against randomized experiments as the supposed only approach to causal inference and has identified alternatives for reaching cause-and-effect conclusions beyond a reasonable doubt. He sees evaluation as a vital transdiscipline that is inherent in all disciplined intellectual and practical endeavors and that needs to be developed and maintained as a discipline in its own right. Most recently, he has written of the evaluation discipline's past, present, and future revolutions.

REVIEW QUESTIONS

1. Without trying to formulate a verbatim response, how would you characterize Scriven's definition of evaluation? What is the role of judgment in Scriven's approach to evaluation?
2. What is Scriven's definition of needs assessment? How does he distinguish this definition from what he identifies as a commonly accepted but flawed concept of needs assessment? What is the role of needs assessment in Scriven's evaluation approach?
3. Why did Scriven distinguish between the goal and roles of evaluation? What are formative and summative evaluations? How are they related to the distinction between goal and roles? How might Scriven and Cronbach have differed in their rating of the relative importance of formative evaluation and summative evaluation?
4. What is the distinction between goals-based and goal-free evaluation? Why did Scriven say that goal-free evaluation is reversible? What general process is involved in conducting a goal-free evaluation? What is the role of needs assessment in a goal-free evaluation?
5. What are the essential meanings of the basic evaluation acts of scoring, ranking, grading, and apportioning? How are they different? How are they complementary?
6. What is metaevaluation? What roles are served by metaevaluation? Who should conduct metaevaluations? What does Scriven see as the relevance of the Key Evaluation Checklist to conducting a metaevaluation?
7. What is involved in the final synthesis step in an evaluation? What has Scriven identified as flaws in evaluators' typical practices in reaching bottom-line evaluative conclusions? How has he addressed these flaws in his recommended approach to synthesizing evaluation findings?

8. Why is Scriven critical of evaluators' use of the numerical weight and sum approach to reaching evaluative conclusions? What does he see as the most grievous mistakes in using this procedure? What is his qualitative weight and sum approach? What reasons does he give for using the QWS procedure to overcome the mistakes involved in applying the NWS approach?
9. Why does Scriven reject the position that randomized controlled experimental design is the gold standard for reaching cause-and-effect conclusions? What general concept does he advocate for assessing a program's causal connections to outcomes? What does he see as potentially acceptable alternatives to the experimental design approach to establishing cause-and-effect relationships?
10. What is the essence of Scriven's characterization of the third revolution in evaluation, and how is it related to the two prior revolutions?

Group Exercises

Exercise 1

Critically appraise the claim that the value-free doctrine of the social sciences provides an essential foundation for objectively evaluating a program.

Exercise 2

Define comparative and noncomparative evaluations, and give an example of each type.

Exercise 3

Are formative evaluation and summative evaluation conceptually and operationally distinct concepts? Why or why not? What illustrations support your group's responses to these questions?

Exercise 4

If an evaluator judges a program at one point in time, does she lose her independence and objectivity in regard to her future evaluations of the program? Why or why not? If yes, must the evaluator terminate her relationship with a program once she has submitted her judgment of it? If not, how can the evaluator avoid being co-opted by a program staff that acquiesces to her initial judgments and recommendations? Does metaevaluation have a role in addressing difficulties in this area? If yes, why? If not, why not?

Exercise 5

Define relativistic and objectivist philosophies of evaluation. List what your group sees as the essential differences between these philosophies. Give examples of how two evaluators might

differentially apply the two philosophies in evaluating the same program. Then discuss how the contents of the final reports for these two evaluations would be likely to differ.

Exercise 6

Suppose your group is designing a course on program evaluation and that you have chosen Scriven's Key Evaluation Checklist as the advance organizer for the course's content. Develop an outline for this course.

Exercise 7

Obtain an issue of *Consumer Reports* for reference in completing this exercise. Then review this chapter's section on product evaluation and use its contents to address the following questions:

1. To what extent does the *Consumer Reports* issue that your group obtained employ scoring, ranking, grading, and apportioning in its product reviews?
2. To what extent are the product reviews vulnerable in regard to Scriven's admonitions against using the numerical weight and sum approach?
3. Is this issue of the magazine vulnerable to criticisms of its objectivity and independence? Why or why not?

Notes

1. In adapting the checklist to conduct a particular metaevaluation, presumably one would need to respond to the questions in the checklist, assess client and consumer needs in regard to the evaluation, use this information to determine weights for each criterion in the checklist, and define judgment levels. Then one could use the checklist to collect information about the completed evaluation on each defined criterion, grade it on each criterion, profile it in regard to the criteria, consider its strengths and weaknesses in consideration of the preassigned weights, and reach an overall conclusion about the merit and worth of the evaluation.
2. Scriven has promoted this concept strongly since about 1991. A transdiscipline comprises a number of autonomous applied fields, together with a core discipline whose principal mission is developing tools for use by practitioners in the applied fields and other disciplines. Statistics, measurement, logic, and now evaluation are perhaps the most important examples.

Suggested Supplemental Readings

Alkin, M. C. (Ed.). (2004). *Evaluation roots: Tracing theorists' views and influences*. Thousand Oaks, CA: Sage.

Alkin, M. C. (Ed.). (2013). *Evaluation roots: A wider perspective of theorists' views and influences* (2nd ed.). Thousand Oaks, CA: Sage.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on training* (pp. 171–246). Skokie, IL: Rand McNally.

Coryn, C.L.S., Stufflebeam, D. L., Davidson, E. J., & Scriven, M. (2010). The Interdisciplinary Ph.D. in Evaluation: Reflections on its development and first seven years. *Journal of MultiDisciplinary Evaluation*, *6*(13), 118–129.

Cronbach, L. J. (1963). Course improvement through evaluation. *Teachers College Record*, *64*, 672–683.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*, 116–127.

Donaldson, S. I. (2013). *The future of evaluation in society: A tribute to Michael Scriven*. Charlotte, NC: Information Age.

Fisher, R. A. (1951). *The design of experiments* (6th ed.). New York, NY: Hafner.

Joint Committee on Standards for Educational Evaluation. (1981). *Standards for evaluations of educational programs, projects, and materials*. New York, NY: McGraw-Hill.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39–83). Skokie, IL: Rand McNally.

Scriven, M. (1969). An introduction to meta-evaluation. *Educational Products Report*, *2*(5), 36–38.

Scriven, M. (1969). *Evaluation skills* (Audiotape No. 6B). Washington, DC: American Educational Research Association.

Scriven, M. (1974). Pros and cons about goal-free evaluation. *Evaluation Comment*, *3*, 1–4.

Scriven, M. (1975). *Evaluation bias and its control* (Occasional Paper Series, Paper #4). Kalamazoo: Western Michigan University, Evaluation Center.

Scriven, M. (1983). Evaluation ideologies. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and social services evaluation* (pp. 229–260). Norwell, MA: Kluwer.

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

Scriven, M. (1993). *Hard-won lessons in program evaluation*. New Directions for Program Evaluation, no. 58. San Francisco, CA: Jossey-Bass.

Scriven, M. (1994). The final synthesis. *Evaluation Practice*, *15*, 367–382.

Scriven, M. (1994). Product evaluation: The state of the art. *Evaluation Practice*, *15*, 45–62.

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

Scriven, M. (2005). Causation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 43–47). Thousand Oaks, CA: Sage.

Scriven, M. (2005). Checklists. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 53–59). Thousand Oaks, CA: Sage.

Scriven, M. (2004). Reflections. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 183–195). Thousand Oaks, CA: Sage.

Scriven, M. (2011). *Conceptual revolutions in evaluation: Past, present, and future*. Unpublished manuscript.

Scriven, M. (2011). *Evaluating evaluations: A meta-evaluation checklist*. Claremont, CA: Claremont Graduate University.

Scriven, M. (2011). Evaluation bias and its control*. *Journal of MultiDisciplinary Evaluation*, *7*(15), 79–98.

Scriven, M. (2011). The Faster Forward Fund. *Journal of MultiDisciplinary Evaluation*, *7*(15), 313–317.

Scriven, M., & Roth, J. (1990). Special feature: Needs assessment. *Evaluation Practice*, *11*, 135–144.

CHAPTER 15

ROBERT STAKE'S RESPONSIVE OR STAKEHOLDER-CENTERED EVALUATION APPROACH

Stake is the leading theorist in the school of evaluation that we have categorized as social agenda and advocacy evaluation.[1] He has contributed uniquely to evaluation's philosophical and theoretical development. In response to the sweeping federal requirements for evaluation that were imposed on American education in the 1960s, Stake (1967) introduced a new approach that became known as the countenance model for evaluation. This approach built on theorist Ralph W. Tyler's notion (1942) that evaluators should compare intended and observed outcomes, but it broadened the concept of evaluation by calling for examination of background, processes, standards, and judgments as well as outcomes. Stake developed his philosophy of evaluation during the late 1960s and early 1970s, and in 1975 (1975a, 1975b) he presented his extended view under the label of "responsive evaluation." This presentation retained the countenance approach's emphasis on examining the full countenance of a program, but it broke sharply from the Tylerian tradition of gathering data to discuss whether intentions have been realized. Instead, Stake assumed, in presenting responsive evaluation, that a program's intentions would change and called for continuing communication between evaluator and audience for the purposes of discovering, investigating, and addressing issues. In his early writing, Stake (1975b) stated that a rational judgment in evaluation is a decision as to the relative importance of the standards of different reference groups in deciding whether to take some administrative action. Because of Stake's emphasis

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **Key elements of Robert Stake's professional background and influences**
- **Why Stake became disenchanted with the use of standardized achievement tests for evaluating educational programs**
- **Why and how he advised evaluators and educators to plan and conduct evaluations that take evaluation's full countenance into account**
- **Definitions of the countenance evaluation approach's key components**
- **Details of Stake's reconceptualization and expansion of the countenance evaluation approach into his subsequent approach—referred to here as the responsive or stakeholder-centered evaluation approach**
- **How Stake contrasted responsive evaluation with preordinate evaluation**
- **The main tasks of responsive evaluation, including heavy emphasis on observation of and ongoing exchanges with a program's full range of stakeholders**

on involving and serving the full range of a program's stakeholders (Stake, 1975a, 1975b), we have labeled his evaluation approach responsive or "stakeholder-centered" evaluation.

We see Stake as the leader of the social agenda and advocacy school of evaluation, which calls for a pluralistic, flexible, interactive, holistic, subjective, constructivist, and service-oriented approach. Stake's approach is relativistic in that the evaluator seeks no final authoritative conclusion, instead interpreting findings against the different and often conflicting values of stakeholders. Moreover, Stake has emphasized that the evaluator's judgment, in regard to not only evaluation design but also the expression of the quality of the evaluand, is one of the most important judgments in his approach to gathering and analyzing a wide range of judgments. It is noteworthy that Stake's writings (1967, 1969, 1971, 1974) on evaluation initially focused on education and later were expanded to assist with evaluations in additional disciplines and service areas.

Stake's Professional Background

In the 1950s Stake taught mathematics at the U.S. Naval Academy Preparatory School and later completed a PhD program in psychometrics at Princeton University. In 1963 he joined the faculty at the University of Illinois, where he taught in the educational psychology department and served as associate director of the Center for Instructional Research and Curriculum Evaluation (CIRCE) under Thomas Hastings. Hastings had brought him to Illinois to do research on instruction, but Stake's interest was soon captured by the new work Hastings and Lee Cronbach were doing in curriculum evaluation. When Hastings retired in 1978, Stake became director of CIRCE and held that position until he retired in the early 2000s.

Factors Influencing Stake's Development of Evaluation Theory

Stake's thinking seems to have been influenced by several noteworthy factors. His early training and experiences in mathematics, statistics, and measurement made him conversant with the application of concepts and methods in these areas to the practice of educational evaluation. As he became increasingly skeptical about the classical conception of measurement and its employment in evaluation, his status as a trained expert in these areas gave credibility to his attacks and counterproposals and influenced his audiences to seriously consider his views.

In the mid-1960s he attacked the classical view of evaluation as narrow, mechanistic, and not helpful (Stake, 1967). His disenchantment seems to have been inspired by Cronbach, who until 1964 was a professor at the University of Illinois. Cronbach (1963) had argued that evaluation's basic function in education should be to guide curriculum improvement, not to judge completed, packaged curricula; he had argued further that comparative evaluations of curriculum alternatives based on average posttest test scores were neither informative nor helpful. Stake (1967) later built on these claims when he argued against comparative experiments, called for full descriptions of programs, and emphasized the importance of subjective information.

The influence of R. W. Tyler (1942), who had defined evaluation as determining the extent to which valued objectives have been achieved, was also evident in Stake's early writing.

In his "countenance" paper, discussed further in the next section, Stake (1967) advocated comparing intended and actual outcomes, but also recommended that evaluators consider assessing antecedent conditions and ongoing transactions, intended and actual. In other words, he expanded on Tyler's thinking. This link between Stake's and Tyler's work seems understandable, given that Tyler's conceptualization had been the dominant view of evaluation since the 1940s. (They were also fellow Nebraskans.) Further, both Cronbach and Hastings, who had significant influence on Stake's professional development, had studied under Tyler. In particular, Hastings's research demonstrated to Stake that teachers had little use for the measurements and measurement concepts championed by specialists in educational testing.

Stake also was obviously influenced by Scriven's argument (1967) that evaluators must judge. Stake agreed that judgments must be included in evaluations, but he maintained, for a number of reasons, that evaluators should collect, process, and report other persons' judgments and consider these along with their own.

Another main factor that clearly influenced Stake's views about evaluation was CIRCE's involvement with various federally funded projects in the late 1960s, most of them housed at universities. The projects were developmental in nature; although they were open to study, observation, and feedback for improvement by evaluators, they were neither stabilized nor available for controlled, variable-manipulating investigation by researchers. Many of these projects were educational enrichment opportunities for the gifted or curriculum development institutes for teachers. The federal evaluation requirements were in essence Tylerian, calling for evidence that sponsored projects had achieved their objectives as measured by appropriate achievement tests. Stake and his colleagues judged that available published achievement tests were largely inappropriate for evaluating the federal education projects, especially because published tests did not assess much of what teachers actually taught.

Stake's 1967 "Countenance of Educational Evaluation" Article

Stake's 1967 article, "The Countenance of Educational Evaluation," was offered not as a specific guide for designing an evaluation, but as general background reading for those facing such a task. With this article, Stake wanted to help projects meet the federal evaluation requirements in a manner both acceptable to the government and useful to staffs and other constituencies. In presenting this article, Stake did not intend to offer a model (although many readers perceived his recommendations as constituting one, hence the frequently used term *countenance model*). Instead, he intended to provide an overview of evaluation. By the term *countenance,* he meant the face of evaluation, the whole picture, an overview. He thought that different models (or persuasions) would fit here or there, and the countenance representation was a grid or map on which to locate them. Stake's approach reflects an attempt to adapt and expand on Tylerian evaluation to meet needs current at that time, and it presents a broad view of the many forms of information that can be used to answer the questions of various clients. The article's main purpose was to help readers see the wide range of data that might be used in an evaluation. Following its publication, the article was widely referenced in discussions of educational evaluation.

Stake chose the title of his 1967 article to convey a particular message to evaluation specialists and educators in general. He said that few educators saw education "in the round" (p. 523). In particular, he noted that formal evaluations often focus narrowly on a few variables in a program (such as outcomes associated with objectives), and that informal evaluations often reflect a few people's opinions (but not carefully collected empirical data). He urged educators and evaluators alike to recognize the shortcomings of their usual evaluation practices and forthwith to pay attention to the full countenance of procedures and processes in a sound evaluation and correspondingly to the full countenance of the program being evaluated. The countenance of a sound evaluation, he said, includes (1) description and judgment of a program; (2) data pertaining to the program's intended and observed antecedents, transactions, and outcomes; (3) the program's rationale; (4) analyses of congruence and contingencies; (5) identification of pertinent, often conflicting standards and judgments; (6) a variety of evaluation tasks and associated procedures; and (7) formative and summative uses of findings.

Description

In considering description as a basic act of evaluation, Stake (1967) referenced the prior works of R. W. Tyler (1942) and Cronbach (1963). The proponents of Tyler's approach had focused their descriptive efforts on discerning the extent to which objectives had been achieved. Against the advice of Tyler, who advocated using a wide range of data, they had narrowed their purview by focusing on specific behavioral objectives and employing mainly standardized achievement tests. Stake criticized this narrowness and supported Cronbach's suggestion that educators broaden their concept of achievements and ways of measuring them. Stake (1967) advised educators to implore "measurement specialists to develop a methodology that reflects the fullness, the complexity, and the importance of their programs" (p. 524). More specifically, he stated,

> The traditional concern of educational-measurement specialists for reliability of individual-student scores and predictive validity . . . is a questionable resource. For evaluation of curricula, attention to the individual differences among students should give way to attention to the contingencies among background conditions, classroom activities, and scholastic outcomes. (p. 524)

As discussed later in this chapter, Stake charged educators to fully describe intended and actual antecedent conditions, instructional transactions, and outcomes, and to examine the congruencies and contingencies among them.

Judgment

Stake (1967) agreed with Scriven's position (1967) that an evaluation has not taken place until a judgment has been made. But he questioned the wisdom of assigning the responsibility of judgment solely to evaluation specialists. Giving evaluation specialists sole responsibility for making a judgment he said, would be unrealistic for three main reasons. Educators, perceiving that an evaluator would be the sole judge of their program, would be unlikely to cooperate with

the evaluator's data collection efforts. Moreover, evaluators might be censored or criticized by those among their colleagues who believed that evaluators, acting as judges—as opposed to objective inquirers—would make social science and behavioral research suspect. Finally, Stake suggested that few evaluators would feel qualified to discuss what is best for a briefly known school and community.

To respond to this dilemma, Stake (1967) put forward a compromise position. Although he doubted that evaluators could or should act as sole or final judges of most programs they evaluated, he thought they were uniquely qualified to collect and objectively process the opinions and judgments of other people. His recommendation was that evaluations of school programs should portray the merits and faults perceived by well-identified groups, and he mentioned five groups as having important opinions about education: spokespersons for society at large, subject matter experts, teachers, parents, and students.

This compromise recommendation satisfied Stake's worry (1967) about Scriven's assertion (1967) that evaluators must render the final judgment. Especially, Stake claimed that his recommendation obviated what he believed to be two questionable assumptions underlying Scriven's position on the evaluator's responsibility to judge: (1) that a program judged best would be best for all students, and (2) that local authority to judge is invalid if it results in a judgment that is at odds with the common good. Evaluators would not have to make either assumption if they gathered, processed, and reported judgments from a wide range of a program's stakeholders.

Format for Data Collection

Of central importance to Stake's overall approach are the concepts of antecedents, transactions, and outcomes. Stake (1967) commented that if evaluators would gather information about all of these from a variety of sources, and then analyze and report that information, they would more successfully approach the objective of dealing with the full countenance of evaluation than they would by persisting in their attempts to determine whether objectives had been achieved. Each of these three concepts is complex and requires explanation.

Antecedents

Antecedents refer to relevant background information. In particular, Stake saw this type of information as including any condition existing prior to teaching and learning that may relate to outcomes—for example, whether a student ate a good breakfast before coming to school, whether he completed his homework assignment, or whether he got a good night's sleep; or whether the teachers' union opposed required in-service training participation. Stake argued that to fully describe and judge a program or learning episode, evaluators must identify and analyze the pertinent antecedent conditions.

Transactions

Stake's second class of information, the instructional transactions, includes students' countless encounters with other persons, such as teachers, parents, counselors, tutors, and other

students. Stake advised evaluators to conduct a kind of ongoing process evaluation to discern and document the program's actual operations.

Outcomes

Outcomes pertain to what results from a program. These include abilities, achievements, attitudes, and aspirations. They also include impacts on all participants: teachers, parents, administrators, custodians, students, and others. They include results that are evident and obscure, intended and unintended, short range and long range.

Stake used antecedents, transactions, and outcomes as core concepts to structure his view of what should be done in describing and judging a program. For examining these core program concepts he called for documenting a program's intents and observations—that is, its intended and observed antecedents, transactions, and outcomes.

Intents

By intents, Stake referred to all that is planned for, including antecedent conditions, teaching and learning activities, and desired outcomes; he advised evaluators to study what educators exclude as well as what they include under the heading of "intents" and to express educators' intents in language that is meaningful to them (not necessarily in the form of behavioral objectives).

Observations

Observations refer to antecedents, transactions, and outcomes that are observed and recorded. Consistent with R. W. Tyler's approach (1942), observations may be collected from a variety of sources and using a range of data collection instruments. Stake (1967) advised evaluators to search broadly for the existence of both intended and unintended occurrences.

The following list illustrates the kinds of descriptive information that an evaluator might collect as viewed from Stake's perspective:

1. The teacher said that students would enroll in the music appreciation class because they wanted to be there (intended antecedent).
2. However, 40 percent of the students complained that their parents had coerced them into enrolling (observed antecedent).
3. The music appreciation curriculum guide specified that students were to spend forty minutes a week listening to music and twenty minutes discussing it (intended transactions).
4. The students were observed to spend, on average, nineteen minutes a week listening to music, three minutes discussing it, twenty minutes in a study hall activity, and the remainder of the time doing a variety of other things (observed transactions).
5. At the end of the course, the students were expected, among other things, to be able to name the composers of selected musical pieces played for them (intended outcome).

6. On average, the students correctly named the composers of two out of ten pieces that were played for them; also, unexpectedly, a parent of one of the students contributed a sizable sum of money to help expand the school's music library (observed outcomes).

Although the preceding examples are simplistic, they illustrate one basic message Stake (1967) was conveying: through studying intended and actual antecedents, transactions, and outcomes, evaluators should be stimulated to describe programs more fully than if they zeroed in on outcomes related to objectives. Basically, Stake provided evaluators with a two-by-three description matrix that included intended and observed (on the vertical axis) antecedents, transactions, and outcomes (on the horizontal axis). Stake acknowledged that the boundaries between the cells in this description matrix are vague. But, he said, this situation is unimportant, because the main intent is to stimulate evaluators to think broadly and to give them a heuristic for doing so.

Program Rationale

Stake (1967) directed evaluators, in addition to describing a program of interest in relation to the description matrix, to investigate the program's rationale carefully. In effect, what are the program's philosophical background and purposes? Once informed of the program's rationale, the evaluator can use it as a basis for judging program intents. For example, does the planned program constitute a logical step in implementing basic purposes? Stake also observed that the rationale is of use in choosing reference groups that later will be called on to identify standards and pass judgment.

In concluding his discussion of rationale, Stake cautioned evaluators not to overrationalize a program. Evaluators should avoid imposing their philosophy and logic on the program. They should characterize whatever rationale is found in the language of the program staff, not necessarily their own. Although he did not say so, we presume further that Stake would advise evaluators to call attention to problems they perceive in a program's rationale, such as ambiguity, inconsistency, illegality, or immorality.

Analysis

Following his explanations of description and rationale, Stake (1967) turned to a discussion of ways descriptive information is analyzed. He identified congruence analysis and contingency analysis as the two basic classes of analysis.

Congruence Analysis

Congruence analysis involves asking whether what was intended occurred. Were the observed antecedent conditions congruent with those that were expected? Did teachers carry out the directions of the curriculum guide? Were the intended outcomes achieved, and were there additional outcomes? Congruence analysis essentially is identical to Provus's recommendation (1971) that evaluators search for discrepancies between what was intended and what occurred.

Contingency Analysis

In citing contingency analysis, Stake (1967) argued that because "evaluation is the search for relationships that permit improvement of education, the evaluator's task is one of identifying outcomes that are contingent upon particular antecedent conditions and instructional transactions" (p. 534). He explained further that it is important to investigate contingencies among both intentions and observations. Until he developed responsive evaluation, contingency analysis was Stake's approach to addressing clients' frequent demands for information concerning a program's causes and effects.

The appropriate criterion for identifying and assessing contingencies between intended antecedents and transactions and intended transactions and outcomes is logic. Is it reasonable to assume that the expected background circumstances would permit exercise of the intended instruction, and that the latter would lead to the intended outcomes? Stake observed that evaluators, in conducting logical analyses, must rely on their previous experience with similar populations and programs, suggesting that they might obtain useful insights by studying relevant research literature and, we infer, reports of evaluations of similar programs. Logical analysis of contingencies among intentions is important, as Provus (1971) observed, in guiding judgments about a program's theoretical soundness and structural adequacy.

Contingency analysis of observed conditions, according to Stake, is to be based on the criterion of empirical evidence. Are there correlations between actual background circumstances and observed instructional activities, and between the latter and certain outcomes (unintended and undesired as well as intended and desired)? Can any of the correlations be defended as causal? Stake noted that contingency analyses require data from within the program under investigation and might involve review of data reported in relevant research reports.

Stake also concluded that the requirements associated with contingency analysis necessitate special qualifications for the evaluators of given programs, including familiarity with the relevant theoretical and research literature and prior experience in studying similar programs. Because a single evaluator is unlikely to have these specific qualifications, as well as all the analytical, technical, communication, political, and administrative skills required in evaluation, Stake argued that sound evaluation usually requires a team approach.

Standards and Judgments

The collection and analysis of descriptive and judgmental information and the description of the program's rationale provide the basis for the fourth major feature of the countenance approach: identifying standards and formulating judgments about the program's merit. Basically, Stake (1967) advised evaluators to determine both standards and judgments for the three core concepts of antecedents, transactions, and outcomes.

Standards

Stake (1967) defined standards as explicit criteria for assessing the excellence of an educational offering. He observed that school grades, standardized test scores, and opinions of teachers are not good indicators of the excellence of students; in general, according to Stake, the evaluations

then in vogue did not have wide reference value. He cautioned that in a healthy society, different parties should be expected to have different standards. He also cited and supported the claim by Clark and Guba (1965) that different stages in curriculum development involve different criteria. In regard to the complexity of settling on appropriate evaluative criteria, he advised evaluators to make known, with as much scope and clarity as possible, which standards are held by whom and to take into account both general, pervasive standards and the judgments made by individuals and groups about a particular program.

Judgments

Stake's concept of judgments is inextricably tied to his view of standards. He said, "Rational judgment in educational evaluation is a decision as to how much to pay attention to the standards of each reference group (point of view) in deciding whether or not to take some administrative action" (Stake, 1967, p. 536). Moreover, he identified two types of standards to serve as bases for judgments: absolute standards (personal convictions about what is good and desirable in a program) and relative standards (characteristics of alternative programs that are deemed to be satisfactory).

Evaluation Tasks

Although Stake (1967) has not portrayed evaluation as any kind of orderly process, the following tasks are more or less inherent in the process he recommended:

1. The evaluator collects and analyzes the descriptive information and describes the program's rationale.
2. The evaluator identifies the absolute standards—those formal and informal convictions held by relevant reference groups concerning what standards of excellence the program should meet.
3. The evaluator gathers descriptive data from other programs and derives relative standards against which to compare the program of interest.
4. The evaluator assesses the extent to which the program being evaluated meets the absolute and relative standards.
5. Singly or in collaboration with others, the evaluator judges the program (that is, decides which standards to heed). More specifically, he or she assigns a weight, a level of importance, to each set of standards.

Formative Versus Summative Evaluation

In contrasting relative and absolute standards, Stake (1967) cited a pertinent disagreement between Scriven and Cronbach. Cronbach (1963) had charged that curricula-comparing studies are poor investments because they do not generalize well to the local situation and because alternative programs have evolved to serve the needs of different groups and have different purposes. In general, he had advised evaluators not to conduct comparative studies, but

instead to perform in-depth process studies aimed at helping to improve individual programs. Scriven (1967), in contrast, had called for direct comparison of a program with its "critical competitors" as the best basis for judging a program's merit. Whereas Scriven's favored frame of reference was *Consumer Reports* magazine, which evaluates competing consumer products and services, Cronbach (1963) recognized that the educators of the 1960s needed evaluations not of completed educational products and services, but of the processes involved in developing those products and services. Scriven (1967) acknowledged the need for process studies of the type called for by Cronbach in 1963 (Scriven called these "formative evaluations") but said they were (at least then) of secondary importance in relation to comparative studies aimed at judging a program's relative merit (he called these "summative evaluations").

Stake (1967) saw a need for both types of studies and observed that their relative importance would vary according to the purpose of the evaluation to be undertaken. That is, he argued that the full countenance of evaluation involves different uses of evaluation reports. He saw comparative summative evaluations as needed by an educator faced with a decision of which program to adopt, but not by the curriculum specialist faced with the task of designing and developing a new program or with responsibility for improving an existing program. The latter's evaluation needs would be served better by the formative type of study advocated by Cronbach (1963).

Stake's Advice for Applying the Countenance Approach

In concluding the countenance article, Stake (1967) acknowledged the difficulty of using his approach strictly in accordance with his explanation of it. He reiterated that a team approach is usually required. Specializations to be reflected by the team might include instructional technology, psychometric testing and scaling, research design and analysis, dissemination of information, social anthropology, economics, and philosophy. He also called for the development of new and better ways of processing judgments and for ways of making evaluations less intrusive. In regard to the last point, he said that the countenance of evaluation should be one of data gathering that leads to decision making, not to troublemaking.

In spite of the difficulties of implementing this approach, Stake urged educators to make their evaluations more deliberate and formal. He suggested they clarify their responsibilities in regard to individual evaluations by answering five questions:

1. Is the evaluation to be descriptive, judgmental, or both?
2. Is the evaluation to emphasize antecedents, transactions, outcomes, and/or their functional contingencies?
3. Is the evaluation to emphasize congruence?
4. Is the evaluation to focus on a single program, or will it be comparative?
5. Is the evaluation intended to guide development, or will it be used to choose among available curricula?

Finally, Stake looked to the future and urged that evaluations be used to develop a base of knowledge about education. He urged educators to develop data banks documenting causes

and effects, congruencies among intents and accomplishments, and a panorama of judgments from those concerned with the programs evaluated.

Summary of Key Points Related to Stake's Countenance Approach

The following major points are drawn from the countenance article (Stake, 1967):

- Evaluations should help audiences see and improve what they are doing.
- Evaluators should describe programs in relation to antecedents and transactions as well as outcomes.
- Side effects and incidental gains as well as intended outcomes should be studied.
- Evaluators should avoid rendering final, summative conclusions and instead collect, analyze, and reflect the judgments of a wide range of people having interest in the object of the evaluation.
- Experiments and standardized tests are often inappropriate or insufficient to meet the purposes of an evaluation and should frequently be replaced with or supplemented by a variety of methods, including those that are subjective and may employ soft data as well hard data.

Responsive Evaluation Approach

We turn now to Stake's extension of his philosophy of evaluation, leading to what we know as responsive evaluation. This extension appeared in *Program Evaluation: Particularly Responsive Evaluation,* which Stake (1975b) presented at the Conference on New Trends in Evaluation in Gotenborg, Sweden, and in several other publications (for example, Stake, 2004a). With the issuance of this paper, *responsive evaluation* replaced *countenance evaluation* as the popular term to describe Stake's approach. However, this new formulation retained much of what Stake had included in his 1967 countenance article. The major departure was that he turned sharply away from R. W. Tyler's objectives-based (1942) orientation. In fact, he presented responsive evaluation as much in terms of its differences from Tylerian evaluation as in terms of its own unity, wholeness, and integrity. In essence, then, the 1967 countenance article served as a bridge between Tylerian, or preordinate, evaluation and a new and relatively distinct view of evaluation: responsive evaluation. Stake (1975a) said that in fact, the responsive view reflects the long-standing practice of informal, intuitive evaluation, somewhat formalized.

In introducing responsive evaluation, Stake (1975b) noted that his attention was on evaluation of programs, which might be strictly or loosely defined; big or small; and, we infer, ongoing or ad hoc. He asked his audience, in considering his approach, to assume that someone had been commissioned to evaluate a program and that the program most likely was under way. He noted that there would be a specific client group or several audiences to be served and that usually they would include those responsible for carrying out the program. The evaluator, Stake observed, would be responsible for communicating with the specific audiences. These guiding assumptions clearly are consistent with tenets of the 1967 countenance article:

that the evaluator's point of entry usually comes sometime after a program has started and that the evaluator's main role usually will be to provide useful evaluative information to those persons who are operating the program being evaluated.

Responsive Evaluation Contrasted with Other Approaches

Stake (1975a) identified responsive evaluation as an alternative to eight other evaluation approaches: (1) the pretest-posttest model, preferred by most researchers; (2) the accreditation model involving a self-study and visit by outside experts (liked by educators, according to Stake, if they can choose the visitors, but disliked by researchers because it relies on secondhand information); (3) the applied research on instruction model, advocated by Cronbach (1963); (4) consumer-oriented evaluation, recommended by Scriven (1967, 1974); (5) decision- and accountability-oriented evaluation, proposed by Stufflebeam (1971a); (6) metaevaluation, introduced by Scriven (1969b); (7) goal-free evaluation, conceptualized by Scriven (1973); and (8) adversarial evaluation, advocated by Owens (1973) and Wolf (1975). Stake saw the first two of these as the primary models of program evaluation and chose specifically to present responsive evaluation as a clear-cut alternative to the pretest-posttest model, which he labeled "preordinate evaluation."

Key Proponents of Responsive Evaluation

Stake identified responsive evaluation as an approach being advocated by Parlett and Hamilton (1972), MacDonald (1975), L. M. Smith and Pohland (1974), Rippey (1973), and Abma (2006). Fundamentally, he said, the approach is focused on settings where learning occurs, teaching transactions, judgment data, holistic reporting, and giving assistance to educators. Although Abma (2006) has, as already noted, been a proponent of responsive evaluation, Stake (2013) rejected Abma's view that responsive evaluation should be oriented to action and instead stated that it is acceptable for responsive evaluations to help stakeholders improve their perceptions of an object's quality.

In grounding the responsive approach, Stake (1975a) subscribed to a generalized definition of evaluation that he attributed to Scriven. According to our understanding of this definition, for Stake an evaluation is an observed value of a program or other evaluand compared to some standard for the evaluand. Stake characterized this definition in the following equation:

$$\text{Evaluation} = \frac{\text{Whole constellation of values held for a program}}{\text{Complex of expectations and criteria that different people have for a program}}$$

Stake noted that the evaluator's basic task is neither to solve the equation numerically nor, as he said Scriven had advocated, to obtain a descriptive summary grade for the subject program. Instead, Stake advised the evaluator to make a comprehensive statement of what a program is observed to be and to reference the satisfaction and dissatisfaction that appropriately selected people feel toward the program. In discussing responsive evaluation, Stake did not refer to the evaluator as formally gathering standards, rating them for importance, and reducing the ratings to an overall judgment.

In recent correspondence, Stake noted that what happened—both objectively and subjectively—is the key to describing a program. In regard to the collection, analysis, and reporting of judgments, he noted that these judgments are more intellectual concepts than feelings. Although he said the evaluator should acknowledge feelings, he also stated that these should seldom be the main data.

Responsive Versus Preordinate Evaluation

Throughout his presentation, Stake repeatedly contrasted responsive evaluation with preordinate evaluation. Stake's eleven key distinctions are summarized in Table 15.1.

Purpose

The first and perhaps most telling distinction concerns the inquiry's purpose. The purpose of a preordinate evaluation usually is seen to be focused narrowly on answering a standard

Table 15.1 Main Distinctions Between Preordinate and Responsive Evaluation

Distinction	Preordinate Evaluation	Responsive Evaluation
Purpose	To determine the extent to which goals and objectives were achieved	To help stakeholders discern and address strengths and weaknesses
Scope of services	Meets information requirements as agreed on at the outset of the study	Responds to audience information requirements throughout the study
Written agreements	Obligations of an evaluation's formal parties negotiated and defined as specifically as possible at a study's beginning	Purposes and procedures outlined very generally from an evaluation's outset and evolved during the study
Main orientation	Program intents, indicator variables	Program issues, events
Design	Prespecified	Emergent
Methodology	Reflective of the research model: intervene and observe	Reflective of what people do naturally: observe, interpret, and particularize
Preferred techniques	Experimental design, behavioral objectives, hypotheses, random sampling, standardized tests, summary statistics, and research-type reports	Case study, expressive objectives, purposive sampling, observation, adversarial hearings, expressive reports, and storytelling
Communication between evaluator and client	Formal and infrequent	Informal and continuous
Bases for valuational interpretation	Compares assessed outcomes to prestated objectives, performance by a norm group, or performance of a competitive program	Compares information gathered about a program to different value perspectives of people at hand
Key trade-offs	Sacrifices direct service to those in the program to produce objective research reports	Sacrifices some precision in measurement to increase usefulness
Provisions for reducing bias	Use of objective procedures and an independent perspective	Operational definitions of ambiguous terms and replications of the assessed program, plus comparison of results for the alternative versions of the program

question: To what extent were the preestablished objectives achieved? A responsive evaluation, in contrast, is aimed at helping the client understand problems and uncover strengths and weaknesses in a program (as seen by various groups).

Scope of Services

The second distinction concerns the scope of services that the evaluator provides. In a preordinate evaluation, the evaluator collects, analyzes, and reports findings in accordance with a strict, prespecified plan. The responsive evaluator, in contrast, searches for pertinent issues and questions throughout the study and attempts to respond in a timely manner by collecting and reporting useful information, even if the need for such information had not been anticipated at the study's beginning. In general, the preordinate evaluation's scope is narrow compared with the broad range of issues that might be considered in a responsive evaluation.

Written Agreements

Another distinction is in the formality and specificity of the written agreements governing the evaluation. Often formal obligations of the main parties to the evaluation are agreed to in writing at the study's outset in either type of evaluation. However, contracts for preordinate evaluations are likely to be formal, specific, comprehensive, and binding, whereas those for responsive evaluations are likely to be general, flexible, and open ended.

Main Orientation

A fourth difference between the two types of evaluation is in their respective orientations. Preordinate evaluators examine program intents, including especially the objectives, procedures, and timeline laid out in the program proposal, to decide what information to gather. In effect, they are predisposed to gathering those data required to ascertain whether the program's objectives have been achieved and sometimes whether the program has been carried out as designed. Responsive evaluators do not let the rhetoric of the proposal be so determining. Guided by certain expectations of what will be important, they examine program activities and problems but remain free to settle on certain events or questions as most important. They see preoccupation with program proposals as akin to putting on blinders.

Designs

The two types of studies are guided by designs of considerably different types. Designs for preordinate evaluations are prespecified as much as possible, because the objectives of the study are given and because the controls, interventions, and definitions of constructs common to this type of study need to be arranged at the outset. Designs for responsive evaluations are more open ended and emergent, building to narrative description rather than aggregating measurements over cases. Controls and program interventions are seldom planned in responsive evaluations. The evaluator intends, throughout the study, to discover and respond to those questions deemed important by various stakeholders.

Methodology

Coinciding with the difference in types of designs is a marked difference in the methodological approaches used in the two types of evaluation. Preordinate evaluations in general employ the experimental design approach prevalent in research investigations. Here the evaluator usually intervenes with two or more treatments, assigns them to two different but comparable groups, observes their relative impact on students or other research subjects as measured by a few criterion variables, and tests hypotheses about their differential effects. According to Stake (1975a), preordinate evaluation reflects more a stimulus-response model, and responsive evaluation reverses the sequence. Responsive evaluation is response-stimulus evaluation in the sense that the evaluator responds first—that is, observes a naturally occurring program. The responsive evaluator does not assign subjects to treatments or control the program's delivery. The evaluator does, however, stimulate actions in the program by reporting what has been observed to the client group. In general, the methodologies of preordinate evaluation and responsive evaluation are experimental and naturalistic, respectively.

Preferred Techniques

Different techniques are preferred in the two approaches. In preordinate evaluation, the techniques of choice are experimental design, behavioral objectives, hypotheses, random sampling, standardized tests, inferential statistics, and research-type reports. Techniques preferred by responsive evaluators are case study (Stake, 1988, 1994, 1995; Stake, Easley, & Anastasiou, 1978; Stauffer, 1941); expressive objectives; purposive sampling; observation; adversarial hearings; narrative reports; and storytelling to provide stakeholders with vicarious experiences related to the program (that is, feelings about the program's nature and quality based on imagined participation in the experiences of the program's participants).

Communication Between Evaluator and Client

Communication between evaluator and client in the two types of studies serves different purposes. In a preordinate study, communication is employed to reach advance agreements about how and why the study will be conducted, to check periodically during the study to ensure that participants are fulfilling their responsibilities, and to present the final report. In general, the preordinate evaluator tries to communicate formally and infrequently with the client. Conversely, communication between the responsive evaluator and client is intended to be informal and frequent. As opposed to being prearranged, communication in responsive evaluation should occur more naturally. Stake (1975a) viewed a fairly relaxed and continuous exchange between evaluator and client as essential, because the intent is to carry on a continuous search for key questions and provide the client with useful information as it becomes available.

Bases for Valuational Interpretation

The two types of evaluation also differ in their respective approaches to valuational interpretation. In attaching value meaning to observed outcomes, the preordinate evaluator refers to

prestated objectives or to the performance level of a norm group or students in a competitive program. The responsive evaluator does not necessarily exclude these sources, but he or she is sure to refer to the different value perspectives of those people actually involved in the specific program under study. Moreover, the responsive evaluator does not seek a single conclusion about the program's goodness or badness, but instead tries to reflect all the interpretations obtained from the reference groups.

Key Trade-Offs

Stake consistently acknowledged in his writings (1975a, 1976) that evaluations serve a wide range of purposes and legitimately may follow different approaches. Although he explained his general preference for responsive evaluation, he also noted that there are always trade-offs with any approach. Specifically, he said that preordinate evaluations sacrifice direct service to those in the subject program to produce more rigorous, objective research reports. Responsive evaluations sacrifice some precision in measurement to increase the usefulness of the reports for those involved in the particular program.

Provisions for Reducing Bias

The eleventh and final distinction between preordinate evaluation and responsive evaluation concerns provisions for reducing bias, a dominant theme in preordinate evaluation. In preordinate evaluations, objective procedures and independent perspectives are employed to ensure that the obtained information will stand certain tests of technical adequacy. Responsive evaluation also has tests of technical adequacy, but they are less easily verified. Responsive evaluation emphasizes the importance of subjective information and deemphasizes the use of standardized, objective techniques, allowing some greater bias. Stake (1976) has maintained that there are other ways to reduce bias. In particular, he has urged responsive evaluators to check for the existence of stable and consistent findings by employing redundancy in their data-gathering activities and replicating their case studies. He also advised that they promote understanding of their reports by presenting operational definitions of ambiguous terms. But, all in all, Stake has been more willing to leave bias for the reader to identify and interpret.

Table 15.1 presented our perception of the main distinctions between preordinate and responsive evaluation. Table 15.2 contains a different kind of comparison: Stake's estimates (1975b) of the percentages of time that preordinate and responsive evaluators would devote to different evaluation activities in a typical case. Stake's estimates reflect, so far as we know, only his opinion and may be considerably at odds with other views of reality.

Centrality of Communication in Responsive Evaluations

In discussing his comparison of preordinate evaluation and responsive evaluation, Stake (1975b) emphasized that a main thematic difference is in the purposes, amounts, and kinds of communication with the client. The preordinate evaluator communicates with the client

Table 15.2 Stake's Estimates of How Preordinate and Responsive Evaluators Allocate Their Time

Tasks	Preordinate Evaluation	Responsive Evaluation
Identifying issues, goals	10%	10%
Preparing instruments	30%	15%
Observing the program	5%	30%
Administering tests	10%	—
Gathering judgments	—	15%
Learning client needs	—	5%
Processing formal data	25%	5%
Preparing informal reports	—	10%
Preparing formal reports	20%	10%

Source: Adapted from Stake, R. E. (1975). *Program evaluation, particularly responsive evaluation* (Occasional Paper Series, Paper #5). Kalamazoo: Western Michigan University, Evaluation Center.

before the study to establish the conditions necessary to carry it out; little or not at all during the study because such communication then might bias the way the program operates; and formally at the conclusion of the study through a printed report conveying a detailed description of evaluation design, activities, and findings. In contrast to this characterization of stilted, mainly one-way communication, Stake's depiction of the responsive evaluator is one of someone engaging continuously in two-way communication to learn of important issues that the client wants investigated and, as information becomes available, to provide the client with useful feedback.

Stake (1975b) charged that reporting by preordinate evaluators is too focused and limited. He said that preordinate evaluation's formal and technically sophisticated appearance causes clients to mistake its messages for truth too easily. He also said that because of the preordinate evaluator's dependence on mathematical equations and formalistic prose, she or he is unlikely to tell the client what the program was actually like.

To help them avoid the pitfalls of preordinate reporting, Stake advised responsive evaluators to develop their powers of communication, indicating that they should use whatever techniques are effective in helping their audiences gain a vicarious feel for the nature of the program. He suggested using storytelling (see Denny, 1978) to portray complexity, and said that more ambiguity rather than less might be needed in their reports. In general, he said, evaluation reports should reveal the multiple realities of an educational experience—that is, the different views and understandings that different participants and observers have in regard to the experience.

In rounding out his comparison of the two approaches, Stake said that responsive evaluation will be criticized for its sampling error, but that the size of the error may be small compared with the gains through improved communication with the audience. He acknowledged, however, that the preordinate approach is needed and sometimes does a more effective job.

Substantive Structure of Responsive Evaluation

Beyond contrasting preordinate evaluation and responsive evaluation, Stake (1975b) expanded on his concept of responsive evaluation. He did so especially by describing its substantive structure (we infer that structure to include the types and sources of content to be considered in planning and conducting a responsive evaluation) and functional structure (we infer that structure to include the generic tasks to be completed in conducting a responsive evaluation).

Stake identified advance organizers as the first part of the substantive structure. In responsive evaluation, he saw these as issues. We infer that by the term *issues* he meant areas of disagreement, uncertainty, and concern. He said an issue is a useful advance organizer because it reflects a sense of complexity, immediacy, and valuing. Although it provides direction for investigation, it militates against the narrowly focused gathering of quantitative data. Stake advised the evaluator (1) to become familiar with the program by talking with people, (2) to reach a mutual understanding of the existence of certain issues, and then (3) to use the issues as a structure for further discussions and for developing data collection plans. He emphasized that the evaluator should identify and respond to issues throughout the evaluation.

Stake identified the second part of the substantive structure of responsive evaluation as consisting of the data collection format in his countenance article (Stake, 1967). Looking beyond issues, he saw this format as providing further structure for data gathering. Using the data-gathering format of the countenance model, the evaluator would seek to identify multiple, even contradictory, perspectives and to check on congruencies and contingencies. The relevant observations include the program's rationale; its intended and observed antecedents, transactions, and outcomes; various standards that different groups believe the program should meet; and their different judgments of it.

Stake identified human observers as the third part of the substantive structure of responsive evaluation. He underscored their importance, claiming that they are the best instruments for investigating many evaluation issues.

The fourth and final part of the substantive structure of responsive evaluation is validation. The responsive evaluator, according to Stake, must obtain sufficient information from numerous independent and credible sources to accurately represent stakeholders' perceptions of the program's status, however complex.

Functional Structure of Responsive Evaluation

Stake next considered the functional structure of the responsive evaluation approach. In discussing how to evaluate responsively, he said that the approach requires a large expenditure on observation. Further, he said that there are no linear phases: observation and feedback are important throughout the evaluation.

Twelve Key Tasks in Responsive Evaluations

Having given these provisos, Stake presented the functional structure of responsive evaluation in the form of twelve tasks that might be represented as the hours on a clock. He emphasized

that this is not a standard clock; it moves clockwise, counterclockwise, and cross-clockwise in whatever way is required to best meet the client's needs. We presume that he intended the notion of a clock only as a heuristic, not as a set of technical guidelines, because in the article he neither explained nor illustrated its use. We offer the following interpretation of how an evaluator might address the twelve tasks:

Twelve o'clock: The evaluator talks with the client, program staff, and audiences. These exchanges occur often during the evaluation and touch on a wide range of topics, such as whether the client wants an evaluation and, if so, why; what the client sees as the important questions; what the client thinks of the evaluator's representations of value questions, activities, curriculum content, or student products; and the like.

One o'clock: The evaluator, in collaboration with the client, examines the scope of the program to be evaluated. Often what is inside and outside a program is perceived variously and ambiguously.

Two o'clock: The evaluator overviews program activities. This is a rather unstructured, exploratory, characterizing activity, because the step at seven o'clock calls for structured observations using some of the data collection constructs provided by the countenance article.

Three o'clock: The evaluator seeks to discover purposes for the evaluation and concerns that various people have about the program.

Four o'clock: The evaluator analyzes the issues and concerns and synthesizes them to provide a basis for determining data needs. To accomplish this conceptualization, the evaluator might gather different viewpoints of what is and is not currently worthwhile in the program and what should be added.

Five o'clock: The evaluator identifies data needs with respect to investigating the issues. This would be a rather interactive derivation from the issues' conceptualization, involving working back and forth between potential sources of data and contexts where the issues and concerns are best investigated.

Six o'clock: The evaluator plans the data collection activities, makes a plan of observations, selects observers and instruments (if any), identifies records to be examined, selects samples (perhaps), and arranges for observations and other data collection activities.

Seven o'clock: The evaluator observes antecedents, transactions, and outcomes. We presume the evaluator would also examine the program's rationale and collect standards and judgments pertinent to the program's antecedents, transactions, and outcomes.

Eight o'clock: The evaluator analyzes the obtained information by developing themes seen in the information, using it to prepare portrayals of the program and perhaps doing case studies. With the help of observers, the evaluator might develop brief narratives, product displays, graphs, photographic displays, sketches, a sociodrama, taped presentations, and the like.

Nine o'clock: The evaluator checks the validity of findings and analyses. Various tests of the quality of records are conducted. Program personnel then react to the portrayals, especially in terms of their accuracy and importance.

Ten o'clock: The evaluator winnows and formats information to make it maximally useful to audiences. Audiences should be informed of the assembled data and queried on what information would be of most value to them. Reactions should be collected from authority figures and other audience members. The evaluator should then design forms of communication so as to maximize available information, so that he or she can respond to the different needs of the difference audiences.

Eleven o'clock: The evaluator prepares formal reports if they are required. Depending on prior agreements with the client and audience needs, a printed report may not be necessary.

Responsive Evaluation's Overall Strategy

We suspect that the main message of the evaluation clock is to be found not in its twelve tasks, but in the general strategy it implies. Stake elaborated on this responsive evaluation strategy in terms of its utility and legitimacy compared with that of preordinate evaluation. He observed that explicitness (of referenced information, we assume) is not essential to indicate worth, and that the type of explicitness advocated in preordinate evaluation increases the danger of misstatement (and, we assume, misguided confidence in the findings). In deciding how much and in what form to communicate, according to Stake, audiences' purposes are paramount. He said that different styles of evaluation will serve different purposes and noted that the evaluator may need to discover what bases for evaluative conclusions the members of the audience honor as legitimate. Stake claimed that responsive evaluation can be useful in both formative evaluation (when the staff needs help in monitoring its program) and summative evaluation (when audiences want to understand program activities, strengths, and shortcomings, and when the evaluator feels a vicarious experience should be provided). He acknowledged that preordinate evaluation is the preferred approach when assessments of goal achievement, the keeping of promises, or the verifiability of hypotheses are sought. He also agreed that the measures in preordinate evaluation are more objective and reliable. Nevertheless, Stake concluded that all evaluation should be adaptive; obviously, he saw responsive evaluation as clearly superior in meeting this standard.

An Application of Responsive Evaluation

An early application of responsive evaluation (Stake & Davis, 1999) was an evaluation for the Veterans Benefits Administration (VBA) of its program to improve letters being written by VBA staff to veterans. This program, Reader Focused Writing (RFW), was recommended by the Task Force on Simplified Communication to reengineer all of VBA's written communication. Long-standing dissatisfaction with VBA's written responses to veterans' requests for VBA services stimulated VBA leaders to develop a program to systematically train VBA staff in how

to write effective letters in response to veterans' requests. The program was introduced in 1976 and discontinued in 1997.

The Client's Request for an Evaluation

Following the program's discontinuation, VBA leaders approached Stake to conduct essentially a summative evaluation of RFW. The evaluation was to be conducted within ninety days and to be supported by $30,000. Its goals were to evaluate RFW's focus, impacts, and training methods; to assess VBA's infrastructure and its support of RFW; and also to determine ways to strengthen the program.

First Steps in Planning the Evaluation

In planning the evaluation, Stake and his colleagues decided to give emphasis to regional office staff members who had participated in the satellite RFW training sessions. The evaluation team's first step was to create a list of issues for use in studying the implementation and quality of the training. The identified issues were then used to determine seven essential evaluation questions and nine others that might or might not be important.

Formulating the Evaluation Questions

The evaluation questions initially judged most important focused on RFW's effects in producing better letters to veterans and its practices at the regional offices; quality of the training and of the supervision of writers; trainees' and administrators' impressions of the program; and side effects. Evaluation questions judged of lesser importance focused on such matters as analysis of staffers' varying writing responsibilities, staff turnover, skills needed to understand veterans' needs, interaction of voice and written counseling, advantages and disadvantages of standardizing training, and VBA's support for mainstreaming and continuing to strengthen letter-writing practices.

Next the evaluators engaged stakeholders—sponsors of the training, regional office staffs, trainers, trainees, and the veterans to whom letters were written—to review the draft questions and revise and extend them to ensure the evaluation would address stakeholders' most important questions. Here we see that responsive evaluation is very much an interactive process, with evaluators preparing stimulus materials, stakeholders reacting to the materials, and evaluators responding to best address stakeholders' interests and needs.

Proceeding with a Responsive Evaluation

From this point forward, the evaluation was very much a responsive evaluation. The methods were mainly qualitative, relying heavily on observation, document review, and discussion and little on precise measures and statistical analysis. Stakeholders were engaged throughout the evaluation. The pervasive aim was to understand program implementation, context, and outcomes, with findings to be based heavily on stakeholders' experiences and perceptions of RFW's value. Reporting of findings was intended to be oriented toward enabling readers to reach their

own conclusions about RFW's importance, quality, effectiveness, strengths, and weaknesses. In the end, though, the evaluators synthesized stakeholders' assessments and incorporated them into their bottom-line judgments of RFW.

Substantive Structure

The evaluation's substantive structure included evaluations of RFW's needs, goals, and plan; the impact of the training; the quality of the training approach; the impact of VBA infrastructure on the training program; and grounds for continuing RFW. As in Stake's countenance of evaluation approach, the collection of information included variables related to the program's background and context, transactions, and outcomes, and also its rationale. Short of presenting specific recommendations, the evaluation plan also provided for considering the pros and cons of reinstituting RFW and how this might be pursued.

Functional Structure

The evaluation's functional structure is summarized in Table 15.3. Five training sites were involved. As shown, study responsibilities were divided among four principal evaluation team members. Among the key methods were document analysis (including pre- and posttraining comparison of letter files), characterization of RFW, site visits, telephone and face-to-face interviews, conference calls, focus groups, surveys, letter-writing simulations, and writing of interim and final reports. (It is amazing how much data collection and interaction with stakeholders occurred, given the small evaluation budget and the study's short timeline.)

Metaevaluation

The final report included a metaevaluation chapter by Stephen Kemmis, a professor in Victoria, Australia. Kemmis is an internationally recognized educational researcher. Early in his career, he served on the faculty of the University of Illinois, where he closely worked with Stake. Kemmis conducted the metaevaluation by reviewing the evaluation's archives, interviewing evaluation team members, attending a meeting of the evaluation's advisory committee, and conducting an "evaluation court." He completed the metaevaluation while the final evaluation report was being prepared. Thus, the metaevaluation was a source of formative input for the evaluators and also, perhaps, a working draft of a final summative metaevaluation report, had one been requested. Notably, Kemmis had almost no opportunity to assess the evaluation's impacts on the letter-writing program.

Basically, Kemmis's metaevaluation was goals based, primarily because he assessed the extent to which the evaluation achieved its goals. His metaevaluation also had elements of responsive evaluation in his airing of concerns (about the study's quality, control of bias, operation, issues, context, and utility) for discussion in his evaluation court. Participants in the evaluation court were listed as the evaluation team, a representative of the evaluation's advisory committee, and a representative of the VBA central office staff. Possibly a number of additional stakeholder perspectives might have been represented in a fully responsive evaluation court,

Table 15.3 Functional Structure for Evaluating VBA's Letter-Writing Improvement Program

Tasks	First Evaluator Site 1	Second Evaluator Sites 2 and 4	Third Evaluator Site 5	Fourth Evaluator Site 3
Review the program	X (Describe the rationale, goals, operations)			
Interview administrators	X (Pilot, collect, analyze, write up)			
Portray the program	X (Interview the task force, conference call)			
Survey trainees		X (Pilot, collect, analyze, write up)		
Survey regional directors		X (Pilot, administer, analyze, write up)		
Complete the a follow-up writing task		X (Design, collect, analyze, write up)		
Conduct a pre- and posttraining comparison of letter files			X (Design, collect, analyze, write up)	
Design instruments and protocols			X (Design)	
Direct the evaluation			X (Supervise, coordinate reporting)	
Conduct telephone interviews with veterans				X (Pilot, administer, analyze, write up)

especially representative groups of letter writers and veterans. It is noteworthy that Kemmis focused the metaevaluation on the evaluation team's goals rather than on the full range of published professional standards for evaluations, such as those keyed to an evaluation's utility, feasibility, propriety, accuracy, and evaluation accountability.

Basically, Kemmis's assessment was positive. He credited the evaluators for developing a strong understanding of RFW; producing well-founded findings; and, in the main, guarding against distortion and bias.

He also cited limitations, including concerns about

- The study's quality (limitations of a ninety-day time frame, Kemmis's inability to observe actual training, difficulty in obtaining and analyzing telephone interview information, questions about balance in reporting issues)
- Insufficient attention to assessing the agency's commitment to continuing and funding the program

- The study's operation (Kemmis's report was not clear about what he meant by this concern)
- The report's possibly inadequate reflection of stakeholders' assessments of RFW's conduct and impacts
- The evaluation's context (again, the report wasn't clear about this concern)
- The evaluation's impacts
- The generalizability of findings

This metaevaluation's clear strengths are that it first and foremost addressed the evaluation team's evaluation goals, and provided the readers of the final evaluation report not with definitive metaevaluative conclusions, but with issues to take into account as they considered the evaluation team's description of what the team did and the reported conclusions. Limitations of the metaevaluation are that Kemmis did not have the advantage of looking back on the evaluation's completion and its impacts on client decisions and actions, that in places—as just mentioned—Kemmis's conclusions were cryptic, and that it did not address the full range of standards for a sound program evaluation. However, Stake's inclusion of a metaevaluation—especially given the evaluation's small amount of funds—was a positive aspect of this evaluation of RFW overall.

Reporting

The evaluation's final report was divided into seventeen sections. The first five described the evaluation's foundation: the need for improved letter writing, RFW's history and training period, VBA, issues of RFW training, and the evaluation plan. Sections 6 through 9 covered general classes of obtained information: that gathered from interviews with veterans, views of trainees and administrators, and evidence of improved letter writing. Sections 10 through 14 were reports of visits to RFW sites in Indianapolis, St. Petersburg, Boise, Baltimore, and Denver. Section 15 was the metaevaluation report. Section 16 was a synthesis of findings of RFW's quality. Section 17 provided a summary.

Overall, the report was rich in substance. The evaluation team's report directly addressed the evaluation goals agreed to by the evaluation team and the client. Also, it surfaced and addressed issues beyond the goals.

Curiously, the report had no introduction. Mainly, the introduction to the report was what could be inferred by reviewing the cover and the cover page. Also, the report included no appendix of evaluation materials, no executive summary, and no information on the evaluation team members' qualifications and experience.

Possibly members of the evaluation's audience weren't interested in such additional elements (ones usually included in sound evaluation reports). Maybe they found the report, as presented, sufficient to meet their needs. Because the metaevaluation was embedded in the evaluation process and not a retrospective assessment of the evaluation, the metaevaluator was

unable to assess how members of the report's intended audience judged its quality and used its findings.

Stake's Recent Rethinking of Responsive Evaluation

After reviewing a previous draft of this chapter, in personal correspondence Stake asked us to add up-to-date material to our characterization of his position on responsive evaluation. He stated that he now opposes views (such as those by Abma [2006]) of responsive evaluation as primarily formative and action oriented. Instead, he sees responsive evaluation as ending with perception of quality, not with action. On this point Stake (2013) has counseled against formative evaluations that help a project along the way. Instead, he has advocated finishing an evaluation before designing change. In his 2013 publication he also stressed that informal evaluation and formal evaluation are and should be part of the same act. Further, Stake indicated that he sees all evaluations, regardless of labels for evaluation approaches, as responsive in some way, by virtue of their responsiveness to some form of criteria and standards, history and politics, and people and governance. He also reiterated his commitment to serving stakeholders, noting that responsive evaluation tries to "honor diversity of stakeholders by describing the goodness of the evaluand and indicating separately how well it meets standards abiding in different points of view" (Stake, 2013, p. 193). Notably, Stake (2013) stated that "evaluation cannot happen without standards, but they need be neither explicit nor uniform" (p. 193). Sound responsive evaluation, he said, concentrates on the particular evaluand and offers a vicarious experience that needs, neither in writing nor in reading, the language of cause and effect. He said that such evaluation "pushes people to understand the learning, idle time, and long living that the evaluand does, to evaluate it itself more than for its standing among others" (Stake, 2013, p. 194). In the end, Stake (2013) stated, "The prudent responsive evaluator limits assertions to the goodness and badness of action studied" (p. 196).

Summary

Stake is the leading theorist in the social agenda and advocacy school of evaluation. He is credited with two major contributions to the ongoing development of program evaluation theory: the countenance of evaluation approach and his expansion of this approach, responsive evaluation. Stake was influenced heavily by his involvement in evaluating educational reform projects of the 1960s and 1970s and by the writings of Tyler, Cronbach, Hastings, and Scriven.

In presenting his famous 1967 article on the countenance of educational evaluation, he advised evaluators to turn from highly focused, quantitative, preordinate methodologies (keyed to assessing goal achievement) to a more qualitative approach (keyed to an ongoing, interactive assessment of questions of interest to the full range of stakeholders). He counseled evaluators to expand their inquiries beyond assessing intended outcomes to examining a program's rationale, background, transactions, and full range of outcomes—that is, its full countenance. He also

advised evaluators to value, analyze, and summarize judgments from a program's full range of stakeholders. He called especially for collection of a much broader array of information throughout a program than is obtainable from end-of-course, standardized achievement tests. The credibility of Stake's recommendations, which were then quite radical, was enhanced by his strong background in the measurement-dominated evaluation approach in education that he was calling into question.

In 1975 he presented his landmark paper, titled *Program Evaluation, Particularly Responsive Evaluation* (Stake, 1975b). With the issuance of this paper, *responsive evaluation* replaced *countenance evaluation* as the popular term to describe Stake's approach, although this new formulation retained much of the countenance of evaluation approach. The major departure was in deemphasizing the salience of Tyler's objectives-based orientation. In fact, Stake highlighted responsive evaluation's differences from Tylerian evaluation. Key aspects of responsive evaluation are focusing on issues of importance to the full range of stakeholders, ongoing focusing and structuring of data collection in response to stakeholders' emergent interests and concerns, employing a wide range of subjective as well as objective methods, continually interacting with stakeholders to identify their concerns and supply them with timely feedback, and using their often disparate judgments to report evaluative conclusions that might or might not include contradictory elements. In the main, responsive evaluation emphasizes observation and ongoing exchange with stakeholders. Stake noted, however, that in keying evaluations to stakeholder interests, concerns, and preferences, all evaluations should be adaptive, including accounting for the possibility of honoring stakeholder groups' preference for either a preordinate, Tylerian evaluation or a responsive evaluation.

This last point was evident in Stake's evaluation of the Reader Focused Writing program for the Veterans Benefits Administration reviewed in this chapter. In responding to the client's request for an evaluation of RFW, Stake and his team first and foremost keyed the evaluation to determining whether RFW had achieved its goals. Furthermore, the metaevaluator, Kemmis, keyed his metaevaluation to assessing the subject evaluation's goals. Nevertheless, both Stake's team and Kemmis moved past their initial assessment of goal achievement to identify and examine other relevant issues. This sequence of assessments is reminiscent of advice once tendered by Stake's colleague Hastings. He often advised evaluators first to attend to what the client wants—often assessment of goal achievement—and then to extend the evaluation to identify and address further important matters. Clearly, the correct application of responsive evaluation is a complex matter; it requires effective interaction with clients and other stakeholders and resourcefulness, creativity, and skill in applying relevant data collection and reporting methods.

In his 2013 rendering of responsive evaluation, Stake stressed that it should focus on the particular evaluand, combine informal and formal ways of evaluating, and engage and assist stakeholders such that they develop their own deep understanding of the evaluand's goodness and badness. He also counseled against employing responsive evaluation as a formative process for guiding projects, and against adhering strictly to requirements for independent, objective methods.

REVIEW QUESTIONS

1. What were Stake's expressed reasons for titling his 1967 article "The Countenance of Educational Evaluation?"
2. How did Stake's countenance approach build on and also depart from Tyler's objectives-based approach to evaluation?
3. Using a project with which you are familiar, give examples of Stake's concepts of intended and actual antecedents, transactions, and outcomes. Based on your responses, give examples of contingencies among actual transactions and outcomes.
4. Summarize the importance of side effects in Stake's writings on evaluation, and provide three examples of side effects from one or more studies with which you are familiar.
5. What is Stake's position concerning the role of judgment in evaluation and how best to reach conclusions about a program?
6. Why did Stake, after moving beyond his original countenance of evaluation view of evaluation, characterize his more recent rendition of evaluation as responsive or stakeholder-centered evaluation? In your reply, mention the main tenets of responsive evaluation.
7. What are key distinctions between preordinate evaluation and responsive or stakeholder-centered evaluation?
8. What are the main points of agreement and disagreement between the evaluation philosophies of Stake and Scriven?
9. According to Stake, a preordinate evaluator and a responsive evaluator, in responding to the same evaluation assignment, would each allocate about 10 percent of the evaluation's efforts to identifying issues and goals. Using his estimates, however, the two evaluators would allocate strikingly different amounts of time to preparing instruments and gathering judgments. Characterize these differences and similarities.
10. What does Stake identify as a main disagreement between his concept of responsive evaluation and that of Abma?

Group Exercises

A foundation plans to support and engage a group of low-income families to develop their own houses on a tract of land owned by the foundation. Your group has been asked to advise the foundation about the possibility of employing Stake's responsive or stakeholder-centered approach to evaluate the project. Considering the type of evaluation needed by the foundation, respond to the following questions:

Exercise 1

What is an example of a typical timeline of tasks in conducting a responsive evaluation over a period of approximately two years?

Exercise 2

What are examples of questions that might be included in a responsive evaluation of the self-help housing project for low-income families?

Exercise 3

In general, what types of information are typically collected and reported in responsive evaluations?

Exercise 4

In addition to the evaluators, what persons and groups should be involved, and how should they be involved, in a responsive evaluation?

Exercise 5

How would a responsive evaluation generate conclusions about the success of the self-help housing project? Should the foundation expect to see a single, final judgment of the program? Why or why not?

Exercise 6

What might be included in the table of contents for the final evaluation report?

Exercise 7

Stake does not necessarily advocate the keying of a metaevaluation to professional standards for evaluations. Basing your answer on your reading of this chapter, what might Stake recommend as the criteria for use in metaevaluating the evaluation of the self-help housing project?

Note

1. We wish to express appreciation to Stake for reviewing a previous draft of this chapter and providing helpful feedback. Although we have endeavored to validly represent his views on evaluation, we are responsible for the chapter's contents. This expression of appreciation does not mean that Stake necessarily would agree with or endorse all of the chapter's representations.

Suggested Supplemental Readings

Abma, T. (2006). The practice and politics of responsive evaluation. *American Journal of Evaluation*, *27*, 31–43.

Clark, D. L., & Guba, E. G. (1965, October). *An examination of potential change roles in education*. Paper presented at the Seminar on Innovation in Planning School Curricula, Warrenton, VA.

Cronbach, L. J. (1963). Course improvement through evaluation. *Teachers College Record*, *64*, 672–683.

Cronbach, L. J., & Associates. (1980). *Toward reform of program evaluation*. San Francisco, CA: Jossey-Bass.

Denny, T. (1978). *Storytelling and educational understanding* (Occasional Paper Series, Paper #12). Kalamazoo: Western Michigan University, Evaluation Center.

MacDonald, B. (1975). Evaluation and the control of education. In D. Tawney (Ed.), *Evaluation: The state of the art* (pp. 125–136). London, UK: Schools Council.

Owens, T. (1973). Educational evaluation by adversary proceeding. In E. House (Ed.), *School evaluation: The politics and process* (pp. 295–305). Berkeley, CA: McCutchan.

Parlett, M., & Hamilton, D. (1972). *Evaluation as illumination: A new approach to the study of innovatory programs*. Edinburgh, UK: University of Edinburgh, Centre for Research in the Educational Sciences.

Provus, M. N. (1971). *Discrepancy evaluation*. Berkeley, CA: McCutchan.

Rippey, R. M. (Ed.). (1973). *Studies in transactional evaluation*. Berkeley, CA: McCutchan.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39–83). Skokie, IL: Rand McNally.

Scriven, M. (1969). An introduction to meta-evaluation. *Educational Products Report*, *2*(5), 36–38.

Scriven, M. (1973). Goal-free evaluation. In E. R. House (Ed.), *School evaluation: The politics and process* (pp. 319–328). Berkeley, CA: McCutchan.

Scriven, M. (1975). *Evaluation bias and its control* (Occasional Paper Series, Paper #4). Kalamazoo: Western Michigan University, Evaluation Center.

Scriven, M. (1993). *Hard-won lessons in program evaluation*. New Directions for Program Evaluation, no. 58. San Francisco, CA: Jossey-Bass.

Smith, L. M., & Pohland, P. A. (1974). Educational technology and the rural highlands. In L. M. Smith (Ed.), *Four examples: Economic, anthropological, narrative, and portrayal* (pp. 13–52). Skokie, IL: Rand McNally.

Stake, R. E. (1967). The countenance of educational evaluation. *Teachers College Record*, *68*, 523–540.

Stake, R. E. (1975). *Evaluating the arts in education: A responsive approach*. Columbus, OH: Merrill.

Stake, R. E. (1975). *Program evaluation, particularly responsive evaluation* (Occasional Paper Series, Paper #5). Kalamazoo: Western Michigan University, Evaluation Center.

Stake, R. E. (1976). A theoretical statement of responsive evaluation. *Studies in Educational Evaluation*, *2*, 19–22.

Stake, R. E. (1988). Seeking sweet water. In R. M. Jaeger (Ed.), *Methods for research in education* (pp. 253–300). Washington, DC: American Educational Research Association.

Stake, R. E. (1994). Case studies. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 236–247). Thousand Oaks, CA: Sage.

Stake, R. E. (1995). *The art of case study research*. Thousand Oaks, CA: Sage.

Stake, R. E. (2013). Responsive evaluation IV. In M. C. Alkin (Ed.), *Evaluation roots: A wider perspective of theorists' views and influences* (2nd ed., pp. 189–197). Thousand Oaks, CA: Sage.

Stake, R. E., Davis, R., & Guynn, S. (1997). *Evaluation of Reader Focused Writing for the Veterans Benefits Administration*. Champaign-Urbana: University of Illinois, Center for Instructional Research and Curriculum Evaluation.

Stake, R. E., Easely, J., & Anastasiou, K. (1978). *Case studies in science education*. Washington, DC: National Science Foundation, Directorate for Science Education, Office of Program Integration.

Stauffer, S. (1941). Notes on the case study and the unique case. *Sociometry*, *4*, 349–357.

Stufflebeam, D. L. (1966). A depth study of the evaluation requirement. *Theory into Practice*, *5*, 121–133.

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 in education*. Itasca, IL: Peacock.

Tyler, R. W. (1942). General statement on evaluation. *Journal of Educational Research*, *35*, 492–501.

Wolf, R. L. (1975). Trial by jury: A new evaluation method. *Phi Delta Kappan*, *3*(57), 185–187.

CHAPTER 16

MICHAEL PATTON'S UTILIZATION-FOCUSED EVALUATION

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **The roots of Michael Patton's utilization-focused evaluation (UFE) approach**
- **UFE's eclectic orientation and fourteen premises**
- **Patton's definitions of evaluation and UFE**
- **Personal aspects of UFE, including evaluator and user roles**
- **UFE's attention to values and judgments**
- **UFE's ongoing active-reactive-adaptive negotiation process**
- **UFE's six-part framework for clarifying program goals**
- **UFE's approach to collecting and analyzing information**
- **UFE's framework for reviewing evaluation findings**
- **UFE's approach to reporting findings**
- **Strengths and limitations of UFE**

In Chapter 9, emphasis was given to utilization-focused evaluation—in particular, the contributions made to this approach by Patton (1982, 1984, 1997, 2003, 2008, 2012). However, the concept of usefulness of an evaluation has been present in numerous evaluation approaches and models since the early 1970s. Drawn together by dissatisfaction with classical objectives-based and experimental design evaluation approaches, utilization-focused theorists and decision- and accountability-oriented theorists were closely aligned in terms of the theories they produced. As Alkin (2004) pointed out, "This class of theories is concerned with designing evaluations that are intended to inform decision-making, but it is not their only function to ensure that evaluation results have a direct impact on program decision-making and organizational change" (p. 44). He also stated that "Stufflebeam's evaluation approach engages stakeholders (usually in decision-making positions) in focusing the evaluation and making sure the evaluation addresses their most important questions, providing timely, relevant information to assist decision-making and producing an accountability record" (p. 45). Alkin went on to say that "Stufflebeam is positioned as the first name on the use branch of the theory tree" (p. 45). In effect, some evaluation theorists realized early on that one vital aspect of the success of any evaluation is the extent to which it can help bring about discernible change. Although this chapter centers on Patton's work (1982, 1984, 1997, 2003, 2008, 2012) in utilization-focused evaluation, others who have developed this approach or aspects of it are acknowledged. Because Patton (1997, 2008) and other

adherents of this approach subscribe to the underlying importance of widely accepted standards for evaluation, we have included utilization-focused evaluation as an approach to evaluation that merits readers' close consideration.

Those who follow the utilization-focused pattern are disinclined to be concerned only or mainly with the needs of key decision makers; rather, they stress procedures and processes that enhance the usefulness of an evaluation to a broad array of intended evaluation users. Utilization-focused evaluators believe that only in this way will a study attain a sufficient range of substantial and important impacts. As all experienced evaluators and many evaluation clients are aware, evaluation's history is strewn with neglected reports, even though these may have been formulated painstakingly and correctly. An approach such as utilization-focused evaluation places a heavy professional imposition on evaluators, because it requires evaluators to address and meet the utility standards of the Joint Committee on Standards for Educational Evaluation (1994, 2011) and thereby to make their studies clearly usable and used. Figuring out how this may be achieved is the challenge to utilization-focused evaluators. The crucial point is that *evaluators must determine and focus their studies on intended evaluation uses and produce and report findings that an identified group of intended users can and probably will value and apply to program improvement*. Utilization-focused evaluation is a process for developing an evaluation study in collaboration and negotiation with a targeted group of priority users, selected from a wider set of stakeholders, to focus on and effectively address the intended users' intended uses of the evaluation.

Adherents of Utilization-Focused Evaluation

A number of recognized evaluation authors have endorsed the concept of utilization-focused evaluation, including, for example, Carol Weiss (1972); Marvin Alkin (1995, 2011; Alkin, Daillak, & White, 1979; Alkin & Taut, 2003); Michael Patton (1980); Lee Cronbach and Associates (1980); Daniel Stufflebeam (1966a, 1967; Stufflebeam et al., 1971); and the Joint Committee (1981, 1994, 2011). It is noteworthy that the Joint Committee (2011) included utility as one of five major imperatives for a professionally sound evaluation. In the latest (2011) and previous (1981 and 1994) editions of its standards, the Joint Committee has always listed the utility standards first, to emphasize that an evaluation should not be undertaken if there is no prospect of its findings being used.

From the 1970s onward, Alkin (1995, 2011; Alkin, Daillak, & White, 1979; Alkin & Taut, 2003) stressed that, from an evaluation's outset, the evaluator should endeavor to understand the value system of the subject program's users and then assign priority to users' values for interpreting findings. More recently, Alkin (2004, 2011) advocated the use of interactive sessions to encourage and engage program participants to clarify and apply their values as a basis for interpreting findings and judging program outcomes. Alkin (2004) offered one caveat that has a bearing on his own practice, however, acknowledging that there are conditions under which it is not tenable to engage intended primary users in defining interpretive criteria. Under such circumstances, he prefers to present evaluative data as factually as possible without imposing his or any other party's success criteria.

We also acknowledge that Eisner (1991) placed an implied emphasis on valuing and judging the culture and opinions of the wide array of evaluation users rather than those of only key decision makers. He asked, "What is the value of what is happening?" (p. 171), and answered by stating that valuing is a critical element in the evaluation process. A divergence from the decision-making emphasis was evident as he underlined procedures that are in sympathy with the ultimate aim of the evaluation's appealing to a broad spectrum of identified primary users and other stakeholders. Eisner (1991), like Alkin (2004, 2011), argued that evaluation's purpose is not necessarily to serve a narrow realm of decision making, but more broadly to ensure that an evaluation occurs expeditiously and that it informs a wide range of relevant evaluation uses by a wide range of users. Realizing this demanding goal requires considerable evaluator expertise, sustained consultative work, trust building, and mutual desire by the evaluator and a selected client group to ensure that findings will be responsive, clear, timely, actionable, and acted on.

The central theme of utilization-focused evaluation is its consistent emphasis on focusing, conducting, and reporting on evaluations to help users improve their programs or other enterprises. It should be noted that, in advocating decision- and accountability-oriented evaluation, Stufflebeam (1969, 1971a, 1983, 1985, 2013) has consistently posited that all stakeholders in a program make decisions related to their sphere of program involvement, and that sound evaluation processes can and should target and address pertinent evaluation needs of users at all levels and in all parts of a program (see Chapter 13; see also Stufflebeam et al., 1971).

All theorists of utilization-focused evaluation agree that an evaluation's prospects for utility are enhanced by identifying and involving those who have an immediate stake in the evaluation's findings and who manifest a sincere desire to make appropriate use of those findings. Thus, the evaluator must systematically identify, cultivate, and engage such potential evaluation users. If potentially interested and committed stakeholders are not identified and meaningfully engaged, the whole evaluation exercise may prove futile. As discussion later in this chapter on Patton's approach (1997, 2008, 2010, 2012) to utilization-focused evaluation shows, the evaluator must be closely involved in the process of selecting and orienting a committed group of intended evaluation users; obtaining and sustaining, as feasible, their meaningful involvement in the evaluation's planning, implementation, and reporting stages; and, especially, engaging them to apply the findings for program improvement. Clearly, utilization-focused evaluation gives emphasis at all stages of an evaluation to users' ownership of the evaluation and to their use of findings to improve program operations and outcomes.

Some General Aspects of Patton's Utilization-Focused Evaluation

It is generally acknowledged that Patton (1980, 1982, 1984, 1997, 2003, 2008, 2012) is the most prominent developer of the utilization-focused approach to evaluation. A former professor of sociology at the University of Minnesota and a former president of the American Evaluation Association, Patton currently is an independent evaluator. Since the early 1970s,

he has conducted numerous evaluations, taught many evaluation workshops and courses, and published widely on the theory and practice of evaluation. His passion has been to develop methodology and teach its application so that evaluators will conduct useful evaluations and get the findings used. He steadfastly attempts to show how evaluators, and specifically target audiences, can secure and apply evaluation findings that will make a positive difference in attaining their objectives, which may involve combating such problems as crime, disease, ignorance, inequality, malnutrition, mental anguish, poverty, and unemployment. His writings and teachings have strongly influenced many evaluators and clients first and foremost to gear evaluations toward utility, even above such other crucial criteria as technical adequacy and efficiency.

In his textbook *Utilization-Focused Evaluation,* Patton (1997) stated his developed views about this approach. He offered these definitions of program evaluation in general and utilization-focused evaluation in particular:

> Program evaluation is the systematic collection of information about the activities, characteristics, and outcomes of programs to make judgments about the program, improve program effectiveness and/or inform decisions about future programming. Utilization-focused program evaluation (as opposed to program evaluation in general) is evaluation done for and with specific intended primary users for specific, intended uses. (p. 23)

He considers utilization-focused evaluation to be a process for making decisions about a wide range of idiosyncratic issues in collaboration with an identified group of primary users, focusing on the ways they intend to use the evaluation.

Like all other followers of utilization-focused evaluation, Patton (1997, 2008, 2012) begins with a firm belief that an evaluation must be judged by its use. This places the onus of responsibility on the evaluator to design and process a study such that all activities, from start to finish, will give predominant emphasis to the evaluation's utility. The evaluator must therefore focus on intended uses by those who will use the outcomes of the evaluation. This entails a defined assessment of those who are the primary users of the evaluation and their commitment to the evaluation's use. Such specificity is essential.

Significantly, an evaluator is not a distant judge in this process, but rather a facilitator of judgments that are needed at each stage of a study. The evaluator identifies a group of primary intended users, representative as far as possible of the total cohort of stakeholders, whose values (and not those of the evaluator) will determine the nature of recommendations arising from the evaluation. This selected group will apply evaluation findings and implement any recommendations that are made. It should not be assumed that Patton (1997, 2008, 2012) gives the responsibility for the planning and implementation of the evaluation to this selected group. To the contrary, the evaluator is expected to be fully in control of both planning and facilitating the study (matters further explained in this chapter).

Patton (1997, 2008, 2012) has stressed that program review and reporting of findings should permeate all stages of the evaluation and should encompass the steps of description and analysis, interpretation, judgment, and making recommendations. Along the way,

utilization-focused evaluators should draw from the full range of inquiry and communication methods that are appropriate to the information needs of the intended users.

Intended Users of Utilization-Focused Evaluation

Patton (1997, 2008, 2012) maintains that the group selected to represent the wider cohort of stakeholders should be the actual users of the evaluation. He nominates them as the "intended primary users." Commitment to these users is first in his list of fourteen utilization-focused evaluation premises (outlined later in this chapter).

If an evaluation of a program is to be used, the personal factor, according to Patton (1997, 2008, 2012), is primary. The identification of people who clearly have a stake in the program is essential, as is their concern about the outcomes generated by the study and their commitment to its use. Patton's brand of utilization-focused evaluation differs from other forms of utilization-focused evaluation, because he calls for concentrating evaluation efforts not on all potentially interested stakeholders, but only on a carefully selected subset of the potential users. The evaluation must also be dedicated to the enhancement of a study's outcomes and thus must involve the culture and values of the group selected to represent all stakeholders. Moreover, the utilization-focused evaluator must be strongly involved in persuading intended users to commit to respecting, as appropriate, and applying the findings of the (by now) shared evaluation enterprise. It is no small task to involve intended users in an evaluation study at every stage, from beginning to end, and to convince them that potential changes are to their benefit and that they should willingly contribute to the proposed changes.

The utilization-focused evaluator, therefore, focuses on individuals who are users of the evaluation while also acknowledging other users to be served. This entails careful analysis of the total cohort of stakeholders and identifying which users could best represent the interests of all stakeholders. This selection process is vital to the success of the consequent study. The evaluator must identify representatives of the multiple and varied perspectives of those involved in the program.

Focusing a Utilization-Focused Evaluation

Having selected this special client group, the evaluator engages them to clarify why they need the evaluation; what they hope outcomes might be; how the exercise should be conducted; what part they see themselves playing; what type of reports they envisage; and, finally, how they think the findings should be used to improve the program. During this process, the evaluator acts as a guide or mentor, not as an authoritarian figure or dominating expert. Users' choices are aided by the evaluator's supplying a menu of possible evaluation approaches, methodologies, user participation activities, and types of reports (perhaps both formative and summative, depending on the user group's decisions). Adherents of utilization-focused evaluation are adamant that the more closely evaluation users are involved in the planning and execution of evaluation, the greater their focus on the study, and their feeling of ownership, will be.

The highly personal, dynamic, and situational nature of utilization-focused evaluation underlies Patton's summary (2003) of the working relationship between evaluator and client group, which encapsulates the approach that he has so thoroughly developed and that he espouses:

> In considering the rich and varied menu of evaluation, utilization-focused evaluation can include any evaluative purpose (formative, summative, developmental), any kind of data (quantitative, qualitative, mixed), any kind of design (e.g., naturalistic, experimental) and any kind of focus (processes, outcomes, impacts, costs, and cost-benefit, among many possibilities). (p. 223)

The Personal Factor as Vital to an Evaluation's Success

Patton (1997, 2008) has stressed the central importance of the personal factor in successful evaluations. From the evaluator's perspective, the personal factor involves leadership, enthusiasm, sound advice, listening skills, and respecting all members of the selected representative group. These people will need good information; guidance in decision making; an allaying of concerns about aspects of the study; and a strengthening of the many roles they must play if they, the primary users of the evaluation, are going to see the evaluation as successful. Patton's emphasis on the personal factor is supported by other writers and practitioners. For instance, Cronbach and Associates (1980) stated, "Nothing makes a larger difference in the use of evaluations than the personal factor—the interest of officials in learning from the evaluation and desire of the evaluator to get attention for what he knows" (p. 6).

There is no doubt that the personal factor directs evaluators to specific people, their problems, and their interests. In this way, the likelihood that the evaluation will be used is enhanced.

The Evaluator's Roles

Emphasizing the primacy of the personal factor, Patton dismisses the idea of addressing evaluation reports (if they are produced) to audiences other than immediate users. In line with this approach, Patton (1997, 2008, 2012) deliberately narrows the list of potential users to a manageable number, and they alone are to determine the nature of reports and negotiate other aspects of the study, with the evaluator acting more as a consultant than as an independent, objective evaluator. Thus, the utilization-focused evaluator's role as negotiator is paramount. Depending on the circumstances and concurrence of the primary users, the evaluator might play any of a variety of other roles: trainer, group facilitator, problem solver, diplomat, change agent, measurement expert, experimental design expert, qualitative inquiry expert, content expert, creative consultant, internal colleague, independent auditor, policy analyst, or mediator. However, the evaluator will always negotiate with the primary intended users what roles beyond that of negotiator he or she will play. Moreover, the evaluator will conduct and act on the

negotiations within appropriate ethical bounds and in accordance with the evaluation field's standards and principles.[1]

Utilization-Focused Evaluation and Values and Judgments

Patton (1997, 2008, 2012) and other utilization-focused evaluation practitioners have found that the exploration and development of questions for the evaluation cannot be undertaken in isolation from the values of members of the selected client group. It is their program, and their values must undergird any program examination jointly undertaken by the evaluator and them. Because the purpose of utilization-focused evaluation is to obtain findings that will be used, the evaluator must facilitate intended users' selection of values and judgments and the decisions that arise from these. Utilization-focused evaluators agree with proponents of many other evaluation approaches that, by definition, evaluations are grounded in values. However, instead of imposing external values, the utilization-focused evaluator works with particular intended users (the client, service providers, support staff, and beneficiaries) to determine which values should substantiate and validate the collection and interpretation of the needed information.

Patton (1997, 2008) has observed that evaluation use is too important to allow evaluators to choose the questions and render the judgments, particularly because the users are the ones responsible for making and implementing program decisions. The key principle here is that those responsible for applying an evaluation's findings to program processes should have the authority to decide what values are most appropriate, what questions should be addressed (because these are always value laden), what information is most needed, how the information is best acquired, and what interpretations and decisions should be made. Because of its stress on impact, this position gives preference to decision makers' values and questions over those of the program's wider beneficiaries and other stakeholders.

Theoretically, utilization-focused evaluation is aligned with relativistic and constructivist evaluation approaches in regard to the selection and application of values. In general, utilization-focused evaluation is consistent with Stake's responsive evaluation approach (1983), whereby an evaluator assesses a program relative to the stakeholders' values and judgments rather than independently determined criteria of merit and worth. However, whereas Stake (1983) has followed the postmodern line that dismisses the possibility and desirability of unifying values and judgments, the utilization-focused evaluator seeks consensus on both values and judgments to support and expedite the decision-making process. In this latter respect, utilization-focused evaluation accords with the constructivist approach advocated by Guba and Lincoln (1989). This is true when the utilization-focused evaluator engages a truly representative—though select—group of stakeholders, accords equal influence to each member, assists the group in considering alternative values, and subsequently helps them reach consensus on and apply a set of preferred values. Under these circumstances, utilization-focused evaluation departs from the divergent emphasis in Stake's approach (1983) and invokes the consensus development and the hermeneutic processes advocated by Guba and Lincoln (1989).

Employing Active-Reactive-Adaptive Processes to Negotiate with Users

Patton (1997, 2008) has emphasized that the negotiation process between the evaluator and the client group should progress from the planning stage to completion of the study. Much depends on the circumstances of any particular evaluation: its people, its culture, and its idiosyncrasies. Although accepted standards must always prevail, these must be communicated to the specific evaluation users and adapted for their particular use. Negotiating at every turn strengthens the chances of final outcomes that evaluation users will value and therefore that are very likely to be accepted and acted on. From the utilization-focused perspective, there is not one right way to conduct an evaluation; rather, a design should be developed through negotiation that appeals to users and potential users of the evaluation. Patton (2003) reinforced this concept: "The right way . . . is the way that will be meaningful and useful to the specific evaluators and intended users involved, and finding that way requires interaction, negotiation, and situational analysis" (p. 228).

Patton underlined (1997, 2003, 2008) the full meaning of negotiation in describing his active-reactive-adaptive approach to interactive discussion, advice, and general consultation that continue throughout the study between the utilization-focused evaluator and the select client group, the intended users. It has both descriptive elements centered on decision making and prescriptive elements in regard to the evaluator, who must act and react with advice that adheres to standards, such as the program evaluation standards (Joint Committee, 1994, 2011).

Patton (1997) has stated that utilization-focused evaluators must be active in purposefully identifying intended evaluation users and in formulating with these users questions that will shape a study. Such evaluators are reactive in focusing on the thinking of users and responding to their ideas. This process continues until the completion of the evaluation, with the evaluator always showing the flexibility required to accommodate situational changes. This reactivity leads to the adaptive element that again incorporates flexibility, particularly with respect to evaluation questions and the evaluation design as there is an increase in understanding on the part of both the evaluator and users concerning the study situation and developments. Patton (1997, 2008) insisted that utilization-focused evaluators must immerse themselves in challenges as they arise and become patently responsive to users' wishes and views. As users' reactions and thinking will vary and change in any particular evaluation, the evaluator must be aware that approaches will change with every study. Accordingly, the role and methodology of the evaluation will change as the evaluator employs active-reactive-adaptive processes to consider various options throughout the entire evaluation.

As the evaluator plays the roles of both external expert and creative consultant, her or his skills, knowledge, and ethical values are of paramount importance. Just as the utilization-focused evaluator must be wary not to impose a focus on a predetermined methodological approach, so too must members of the selected user group avoid dogmatically imposing their views concerning the way the evaluation should unfold. Negotiated compromises are made, with the stress always placed on the utility of outcomes. The processes of negotiation and consequent decision making and action provide the user group with valuable experiences,

particularly if these processes reinforce the value inherent in change leading to outcomes that will actually be employed by those who claim increasing ownership of the evaluation and knowledge of its purposes.

Patton's Eclectic Approach

Chapter 9 gave prominence to Patton's eclectic evaluation approach (1997, 2008, 2010). As a pragmatic approach, utilization-focused evaluation entails no particular evaluation model, theory, values, system of criteria and indicators, methods, or procedures. Instead, it is a process designed to help specific users examine the evaluation methods cornucopia and the local situation, then choose the model, methods, values, criteria, indicators, and intended uses that best fit the local situation. The utilization-focused evaluator needs a broad repertoire of evaluation ideas and resources and should follow a flexible, responsive, and creative approach to designing and conducting evaluations. The main point in designing an evaluation is to ensure that the intended users' questions are answered in such a way that they will respect, understand, and apply the findings.

To summarize, utilization-focused evaluators and their specific clients may conduct evaluations that are formative and/or summative, qualitative and/or quantitative, preordinate and/or responsive, and naturalistic and/or experimental. They may choose to investigate and report on any of a wide range of indicators: costs (and benefits), needs, attitudes, processes, outputs, outcomes, and impacts. They may also issue written interim and final reports or engage only in verbal exchanges. Although the approach allows for using any evaluation method that applies to a local need, it tends to be more responsive and interactive than preordinate and independent. It calls for problem solving and a creative process of adapting evaluation procedures to meet the local and specific evaluation needs as they emerge.

Discovering the most functional structure and most appropriate methods for the given circumstances requires technical skill on the part of the evaluator and an incorporation of the wishes of the selected user group as they come to a deeper understanding of the study's direction.

Planning Utilization-Focused Evaluations

In the early planning meetings, the utilization-focused evaluator engages a client group to clarify who will be using the findings, why they need the evaluation, how they intend to apply its findings, how they think it should be conducted, and what values should be invoked. After identifying the intended users, the evaluator stresses to them that the study's purpose must be to give them the information they need to fulfill their objectives. Utilization-focused evaluations are not explicitly intended to address social problems. They have that appearance, however, because utilization-focused evaluation client groups often desire to combat certain social problems. Although this approach is not unique in helping users address social problems, its adherents explicitly justify investments in an evaluation by highlighting its potential utility in helping members of the client group address problems that they judge important.

The evaluator facilitates the users' choices by offering advice about possible purposes and uses of, questions about, and reports for the evaluation. This is done not to supply the choices but to help the client group thoughtfully focus and shape the study based largely on the group's culture. The study is targeted at users, who determine the evaluation's focus, required information, how and when findings must be reported, and how they will be used.

Patton (1997, 2008) has advocated goals-based evaluation and extensive efforts to clarify goals and keep them up to date, seeing goals as constituting one useful way to focus evaluations, though not the only way. He suggested a six-part framework for clarifying program goals and using them in the evaluation:

1. A specific target group of beneficiaries
2. Desired outcomes of the group
3. Indicators of each outcome
4. Targeted performance levels on each indicator (if judged appropriate and desired)
5. A detailed data collection plan keyed to the indicators
6. Specification of how findings will be used

Writing about the six framework components, Patton (1997) noted, "While these are listed in the order in which intended users and staff typically conceptualize them, the conceptualization process is not linear . . . The point is to end up with all elements specified, consistent with each other, and mutually reinforcing" (p. 163).

Patton (1997, 2008) has identified and advocated consideration of several alternative bases for focusing evaluations. One comes from Scriven's recommendation (1991, 1993, 2007) that evaluators not consider goals but instead gather information on a broad range of outcomes and judge whether they meet the assessed needs of targeted beneficiaries. Other named foci are future decisions, critical issues or concerns, stakeholder perspectives, and evaluative questions. Looking beyond these, Patton (1997) presented an extensive list of about fifty ways of focusing evaluations. He argued for careful, resourceful focusing of evaluations, seeing studying the wrong issues as a waste of both intellect and emotion. The message is that putting aside the time to carefully focus an evaluation for maximum utility is most beneficial.

Collecting and Analyzing Information and Reporting Findings

All data collection and analysis methods are acceptable in the utilization-focused program evaluation approach. Utilization-focused evaluation's active-reactive-adaptive and situationally responsive approach ensures that the methodology evolves in response to ongoing deliberations and negotiations between an evaluator and client group and under consideration of contextual dynamics. Different information sources and methods are used to address questions from different perspectives and to cross-check findings. As much as possible, a utilization-focused evaluator puts members of the client group in a primary position to determine evaluation methods so that they can make sure the evaluator addresses their most important questions, correctly places emphasis on collecting the right information, uses techniques they respect, and reports understandable information in a timely fashion. The utilization-focused evaluator

must convince stakeholders of the evaluation's integrity and accuracy as well as facilitate users' knowledge of the findings and the appropriate dissemination of these.

Users' values form the basis for interpreting an evaluation's findings, with the evaluator engaging in as much values clarification as needed to ensure that the evaluative information and interpretation serve the users' purposes. The users are actively involved in interpreting findings. Throughout the evaluation process, the evaluator balances concern for utility with provisions for validity and cost-effectiveness.

Utilization-focused evaluators generally concur that they should help their audiences stand outside the program and gain a better perspective on what is occurring. In addressing this purpose, Patton (1997, 2008) noted that the preparation for review and use of findings should start early in the evaluation. For example, the evaluator can engage the user group to examine a set of simulated findings or give them findings and have them make predictions about their outcomes. Presentation of simulated or actual findings increases users' interest and helps build their readiness to examine and use the ensuing evaluation reports. Moreover, discussions associated with such activities can be invaluable in considering what questions, types of data, analyses, and data displays will be most important toward the evaluation's end. Presentation of simulated or actual data can also be instrumental in training the client group in how to view, assess, and use findings and in making their expectations of the evaluation more realistic. Such activities also provide the evaluator with a means of testing the client group's commitment to using the evaluation findings.

Patton's framework (1997) for reviewing findings has four steps:

1. Description and analysis
2. Interpretation
3. Judgment
4. Making recommendations

Patton (1997) emphasized that the primary intended users should be involved in all four steps. In addition, he stressed that the evaluator must make the findings interesting and easy to grasp. When the utilization-focused evaluator finally produces a report, it should be focused on the most important questions. It should be "arranged, ordered, and organized in some reasonable format that permits decision-makers to detect patterns" (p. 307). Finally, the evaluator should keep the message simple.

Summary of Premises of Utilization-Focused Evaluation

Patton (2003) gave a succinct, fourteen-point summary of the premises underlying his version of utilization-focused evaluation. Following is a brief account of these premises, which in many ways are unequivocal statements not only of what has convinced Patton of the worth of utilization-focused evaluation but also of how a utilization-focused study should progress:

1. Commitment to intended users should be the driving force in an evaluation.
2. Strategizing about intended uses is ongoing and continues from the very beginning of the evaluation.

3. The personal factor contributes significantly to use; it is a psychological imperative.
4. Careful and thoughtful stakeholder analysis should inform identification of primary intended users, taking into account the varied and multiple interests that surround any program, and therefore any evaluation.
5. Evaluations must be focused in some way. Focusing on intended uses by intended users is the most useful way.
6. Focusing on an intended use requires making deliberate and thoughtful choices, including judging merit and/or worth (summative evaluation), improving programs (instrumental use), and generating knowledge (conceptual understanding).
7. Useful evaluations must be designed and adapted based on the situation at hand. Standardized "recipe" approaches will not work.
8. Intended users' commitment to using evaluation findings can be nurtured and enhanced by actively engaging them in making significant decisions about the evaluation.
9. High-quality, not high-quantity, participation is the goal. The quantity of group interaction time can be adversely related to the quality of the process.
10. High-quality involvement of intended users will result in high-quality, useful evaluations.
11. Evaluators have a rightful stake in the evaluations they conduct in that their credibility and integrity are always at risk. To minimize these risks, evaluators must be active-reactive-adaptive.
12. Evaluators committed to enhancing use have a responsibility to train users in evaluation processes and the uses of the information.
13. Use is different from reporting and dissemination. Reporting and dissemination may be means to facilitate use, but they should not be confused with such intended uses as making decisions, improving a program, changing thinking, and generating knowledge.
14. Serious attention to use involves financial and time costs that are far from trivial. The benefits of these costs are manifested in greater use.

Strengths of the Utilization-Focused Evaluation Approach

Clearly, Patton has been a leader in advocating and providing practical guidance for securing utility and powerful impacts of program evaluations. He has effectively built on contributions of other authors who stressed the importance of getting evaluations used and has keyed his approach to meeting the evaluation field's professional standards. He has articulated his utilization-focused evaluation approach by clearly identifying and defining a range of important evaluation concepts and by defining active-reactive-adaptive processes for applying the concepts in real-world settings. He has stressed the central importance of identifying and engaging a select, representative subset of intended users in a process of collaborative inquiry to clarify and effectively address their intended uses of findings. And he has advocated an eclectic approach to selectively and responsively employ the full range of sound evaluation

methods. Moreover, he has stressed that evaluators must take into account an evaluation's environment and the culture of the intended users and, accordingly, engage the full range of selected intended users in an interactive, creative process that evolves in response to their interests and needs.

Overall, and laudably, UFE requires a highly competent, responsive evaluator to engage a representative subset of intended evaluation users to

1. Collaboratively define evaluation purposes that are important to them
2. Learn about, value, and engage meaningfully in an ongoing, responsive evaluation process
3. Embrace and assume ownership of the eventual findings
4. Ultimately use the findings according to their purposes for the evaluation

When these intended features are realized, UFE has to be considered exemplary for its success in producing purposeful evaluation impacts.

Limitations of the Utilization-Focused Evaluation Approach

Patton (1997) pointed to turnover of involved users as the main limitation of utilization-focused evaluation. Involving replacement users may require that the evaluation be renegotiated to sustain or renew the prospects for evaluation impacts. And replacements can also derail or greatly delay the process. Furthermore, it is easy to say that this approach should meet all of the Joint Committee's program evaluation standards (1994, 2011), but hard to see how this can be accomplished with any consistency. The approach seems to be vulnerable to bias and corruption by the user group. After all, those involved are only a subset of the program's stakeholders. The intended user group may not represent all the stakeholders' interests if the evaluator has not been successful in recruiting a representative group and in keeping all group members involved. Nevertheless, this possibly biased group is given much control over what is to be looked at, what questions are addressed, and what information is used to address the questions. Moreover, whatever the group's representativeness, stakeholders with conflicts of interest may inappropriately influence the evaluation, especially if the evaluator is inexperienced and vulnerable to manipulation. The involved and empowered stakeholders may inappropriately limit the evaluation to only a subset of the important questions and pertinent bases for interpretation. It may also be close to impossible to have the user group agree on a sufficient commitment of time, resources, and safeguards to ensure an ethical, valid process of data collection, reporting, and use. In addition, if the utilization-focused evaluator assumes such roles as problem solver, diplomat, change agent, content expert, creative consultant, internal colleague, and mediator—as UFE allows—the evaluation may fail to meet standards and principles of sound evaluation focused on preventing conflict-of-interest issues (given the utilization-focused evaluator's interaction with program stakeholders, for example) and ensuring independent perspectives and impartial, objective reporting.

Utilization-focused evaluators face a dilemma. If they fully empower the select group of users to control an evaluation, that group may better accept and use the findings. However, as we have argued, the interests of other important stakeholders may not be addressed. Further, if the

utilization-focused evaluator insists on compliance with professional standards of evaluation, then the stakeholder group may be unwilling to incur the associated consequences, including, for example, unwelcome findings, transparency of findings, and substantial time and financial commitments. As with all other dilemmas, there seems to be no easy means for utilization-focused evaluators to give users their way and also meet the full range of standards of the evaluation field.

Clearly, effective implementation of this approach requires a highly competent, confident evaluator who can approach any situation flexibly, resourcefully, and creatively without compromising basic professional standards and principles. Strong negotiation skills are essential, and the evaluator must possess expertise in the full range of quantitative and qualitative evaluation methods, strong communication and political skills, working knowledge of all applicable standards for evaluation, and commitment to upholding the standards.

None of this is to deny, however, that the utilization-focused approach to evaluation has significant value; of particular importance is its strong requirement for competence, which places a premium on rigorous, selective, and effective training of those who would conduct utilization-focused evaluations.

Summary

Utilization-focused evaluation is a process through which an evaluator works with the primary intended users of an evaluation to make decisions in designing and conducting the evaluation that will best serve those users. Such decisions pertain to the full range of evaluation tasks, including identification of primary users and specific users, selection of data collection methods, analysis of findings, and formatting and reporting of findings, together with offering follow-up support to ensure that findings are used.

The approach is geared to a psychology of use. Systematic involvement of intended users in the entire evaluation process helps ensure that they will develop ownership of the evaluation process and findings, gain the necessary understanding of the information, and consequently act intelligently based on the findings. The evaluator essentially lays the groundwork for use by engaging the users as partners in all stages of the evaluation process. In the most positive sense of the word, the evaluator "co-opts" the users to participate fully in the evaluation process and its application to program decision making. Those in the selected group are encouraged throughout the evaluation to accept the study as their own. Another positive aspect is that utilization-focused evaluation helps the users ensure that the evaluator will tailor the evaluation services appropriately to their needs, priorities, and agendas. Utilization-focused evaluation strives for symbiosis between evaluator and user.

Although UFE subscribes to the evaluation field's standards and principles, it is vulnerable to charges of failing to meet requirements for such standards as those pertaining to impartial reporting when, for example, the evaluator serves as a consultant to the program being evaluated. Also, although sustained, meaningful participation by all members of the intended user group is fundamental to UFE's achievement of impact, this requirement is often difficult to meet.

REVIEW QUESTIONS

1. How does Patton define evaluation in general and UFE in particular?
2. What are the historical roots of Patton's UFE approach, including the main reason why he developed the approach and influences that helped shape it?
3. UFE is classified as responsive in its approach. Discuss its similarities to and differences from two other approaches considered to be responsive: Guba and Lincoln's constructivist approach and Stake's responsive or stakeholder-centered approach.
4. Why would a utilization-focused evaluator need competence in each of the following skill areas, and how might he or she use each skill area in an actual utilization-focused evaluation: group process, values analysis, professional standards for evaluation, content analysis, and negotiation?
5. List as many of UFE's fourteen premises as you can. Then identify the premise that undergirds UFE's psychology of use.
6. Why does Patton insist that users be involved in goals clarification, and what are the six parts in his framework for clarifying program goals?
7. Explain UFE's approach to reporting evaluation findings, and state whether a utilization-focused evaluation should always culminate in a final printed report.
8. Describe the three elements of Patton's active-reactive-adaptive approach, and explain how they are linked.
9. According to Patton, what is the main limitation of utilization-focused evaluation?
10. Give reasons why the "personal factor" is such a vital component of utilization-focused evaluation, and then characterize UFE's prescribed roles for both the intended users and the evaluator.

Group Exercises

Exercise 1

Following utilization-focused evaluation doctrine, what bases should be used for judging an evaluation?

Exercise 2

How would a utilization-focused evaluator address the issue of values in an evaluation? In discussing this question, explore possible difficulties that could arise for the evaluator.

Exercise 3

Suppose your group has been commissioned to assess the adequacy of a utilization-focused evaluator's proposed panel of primary intended users. Bearing in mind the requirements of

utilization-focused evaluation, what factors would you examine to decide on the soundness of the evaluator's selection of panelists?

Exercise 4

Utilization-focused evaluation supposedly follows and adheres to all Joint Committee standards of utility, feasibility, propriety, and accuracy. Yet how can utilization-focused evaluators meet these standards while acceding to the intended users' desires? For example, the targeted group may want to avoid collecting relevant but potentially embarrassing information—contrary to the standards concerned with evaluator credibility, values identification, valid information, and report timeliness and dissemination—and there are numerous other potential conflicts.

Peruse the Joint Committee's *Program Evaluation Standards* (1994), and you will find other possible discrepancies between the utilization-focused approach to evaluation and the stated demands of the standards. Discuss the ramifications of these potential divergences.

Note

1. On the point of a utilization-focused evaluator's acceptably assuming any of a variety of program-related roles, we find it hard to see how he or she could act, for example, as a consultant who helped shape a program and then produce a report that would not be vulnerable to attack for its lack of independence. Such an attack might reasonably charge that the evaluator judged his or her own work and thus violated the Joint Committee (2011) standard addressing conflicts of interest. We recognize, however, that the 2011 edition of the Joint Committee's *Program Evaluation Standards* excludes standards from previous editions (1981, 1994) that require the evaluator's impartial reporting. In this regard, we see the 2011 Joint Committee standard on external metaevaluation as likely to be very important in utilization-focused evaluations that might prove controversial. We believe this observation is relevant, because UFE is keyed to meeting the Joint Committee's program evaluation standards.

Suggested Supplemental Readings

Alkin, M. C. (Ed.). (2004). *Evaluation roots: Tracing theorists' views and influences*. Thousand Oaks, CA: Sage.

Cronbach, L. J., & Associates. (1980). *Toward reform of program evaluation*. San Francisco, CA: Jossey-Bass.

Eisner, E. (1991). 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 11* (pp. 169–187). Chicago, IL: University of Chicago Press.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Thousand Oaks, CA: Sage.

Joint Committee on Standards for Educational Evaluation. (1981). *Standards for evaluations of educational programs, projects, and materials*. New York, NY: McGraw-Hill.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Joint Committee on Standards for Educational Evaluation. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (1980). *Qualitative evaluation methods*. Thousand Oaks, CA: Sage.

Patton, M. Q. (1982). *Practical evaluation*. Thousand Oaks, CA: Sage.

Patton, M. Q. (1984). An alternative evaluation approach for the problem-solving training program: A utilization-focused evaluation process. *Evaluation and Program Planning*, *7*, 189–192.

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2003). Utilization-focused evaluation. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 223–244). Norwell, MA: Kluwer.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2010). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. New York, NY: Guilford Press.

Patton, M. Q. (2012). *Essentials of utilization-focused evaluation*. Thousand Oaks, CA: Sage.

Stake, R. E. (1983). Program evaluation, particularly responsive evaluation. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and social services evaluation* (pp. 287–310). Norwell, MA: Kluwer.

Stufflebeam, D. L. (1966). A depth study of the evaluation requirement. *Theory into Practice*, *5*, 121–133.

Stufflebeam, D. L. (1967). The use and abuse of evaluation in Title III. *Theory into Practice*, *6*, 126–133.

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.

Weiss, C. H. (1972). *Evaluation*. Englewood Cliffs, NJ: Prentice Hall.

PART FOUR

EVALUATION TASKS, PROCEDURES, AND TOOLS

This part of the book addresses the practical and procedural aspects of sound evaluations. We offer a methodology for general application to all sound evaluation approaches. The discussion includes practical procedures and tools, plus many illustrations of their use. Chapters 17 and 18 give practical advice for identifying and responding to evaluation opportunities. In Chapters 19 through 24, we delve into the methodology of evaluation in a sequence that proceeds through an evaluation's start-up, design, budgeting, contracting, information collection, analysis, synthesis, reporting, and follow-up.

CHAPTER 17

IDENTIFYING AND ASSESSING EVALUATION OPPORTUNITIES

This chapter addresses the practical matters of how to identify, assess, and address a range of different types of evaluation opportunities. A key question on the mind of evaluators, especially beginners, is, *How do I find opportunities to apply my evaluation skills?* Evaluators need to distinguish between evaluation opportunities that are worth pursuing and those that are not. An internal evaluator sometimes will need to determine how to ameliorate the negative aspects of an evaluation assignment that cannot be declined.

In this chapter, we draw from our many years of experience to share lessons we have learned related to identifying and assessing evaluation opportunities. On completing the chapter, readers should have a good notion of how to identify and examine evaluation opportunities. In particular, they should know how to decide judiciously whether an evaluation opportunity is worth pursuing or whether it should be avoided if possible. To address cases where the evaluator cannot reject an evaluation assignment, we offer advice on how to proceed carefully and professionally. We also offer our perspectives on how to derive benefit from attending bidders' conferences.

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **Types of evaluation opportunities, including requests for proposal, requests for quotes or qualifications, internal evaluation assignments, sole-source requests, and evaluator-initiated opportunities**
- **Sources of information about evaluation opportunities, especially the *Commerce Business Daily***
- **Types of evaluation agreements, including contracts, grants, and cooperative agreements**
- **Questions to ask in deciding whether to pursue an evaluation opportunity**
- **How to ameliorate negative aspects of evaluation assignments that cannot be declined**
- **How to derive benefit from and avoid pitfalls in attending bidders' conferences**

Sources of Evaluation Opportunities

Opportunities for conducting evaluations come from five major sources: (1) a published or direct mail request for an evaluation (request for proposal [RFP]), (2) a published request for a quote or qualifications (RFQ) to implement an evaluation whose design is typically given, (3) an

assignment given to an internal evaluator to conduct a particular evaluation, (4) a sole-source request given to a particular (usually well-known) evaluator to conduct a study, or (5) an evaluator-initiated proposal to conduct an evaluation that he or she sees as important.

Evaluation RFPs

It is quite common for funding organizations to publish or mail out RFPs. The issuing organization may be a branch of federal, state, or local government; a charitable foundation; or some other organization. The request may be announced in a publication, such as the *Commerce Business Daily,* in which the particulars of the requested study are summarized and information is given on how to obtain the RFP. Alternatively, the requester may mail the full RFP or information on how to get it to preselected groups or individuals that are seen as qualified potential bidders.

Contents of RFPs

Contents of evaluation RFPs are highly variable. Some contain extremely detailed information on the evaluand and any previous evaluations of it, plus the particulars of the needed evaluation. Other RFPs may be quite general, with an indication that the organization wants the bidders to suggest the needed details and to exercise creativity. Both highly specific and more general evaluation RFPs usually indicate the evaluation's timeline, main questions to be answered, needed information, the required reports, a recommended structure for proposals, the criteria for evaluating proposals, the deadline for submitting a proposal, references to relevant background materials, and the persons who can answer potential bidders' questions. Many RFPs give no indication of the amount of money available to support the evaluation or only a general indication of available funds. Some RFPs stipulate that the agreement will be on a cost-reimbursable basis, whereas others call for a fixed-price agreement. Often a published RFP will note the time and place of a bidders' conference where potential bidders will receive an orientation from the sponsor and be able to ask questions pertaining to the competition.

Some RFPs require a fully developed and detailed proposal for the entire evaluation. Others call for initial proposals to develop an evaluation plan. In the former case, the sponsor will probably choose one contractor to proceed with the entire evaluation on a preordinate basis. In the latter case, the sponsor may fund several bidders to produce competitive evaluation plans. The sponsor would then assess the different plans produced under the initial planning contracts and select one or a combination of plans to guide the long-term evaluation.

Contracts, Grants, and Cooperative Agreements

Another variable in evaluation RFPs concerns the nature of the award. The award typically is a contract that specifies the agreements on how the selected bidder will conduct and report on the evaluation. However, the award may be a grant rather than a contract. Under the terms of a grant, the evaluator is given a sum of money under which he or she will have discretion on what questions to address and how to carry out the evaluation. Here the evaluator is given maximum flexibility and professional discretion and needs to account only for how the

money was spent. Still another type of award is the cooperative agreement. This arrangement requires that the evaluator and the sponsor collaborate in conducting and reporting on the evaluation, with the evaluator needing to consult the sponsor on decisions during the course of the evaluation.

Clearly, a grant or contract is the preferred type of award. They are less prone to conflicts of interest between the evaluator and sponsor and undue influence by the sponsor. They are more in keeping with evaluators' need to maintain an independent perspective and edit their own reports.

Identifying RFPs

To learn about evaluation RFPs, evaluators should monitor the *Commerce Business Daily* and other publications that announce evaluation opportunities. They should also make their evaluation qualifications and interests known to potential sponsors by visiting the organizations and submitting relevant printed materials. Depending on the level of rapport developed with an organization, it can be a good idea to make frequent visits to the organization to keep apprised of evaluation RFPs that are being developed. In regard to the development of relationships with funding organizations, it can be highly advantageous to consult with an organization or provide volunteer services. Such services may include evaluating evaluation proposals, critiquing draft RFPs, or even helping develop evaluation RFPs. Although participation in developing an evaluation RFP is likely to preclude one from responding, there are side benefits of offering such a service. In particular, one often becomes privy to other evaluation RFPs that are in the pipeline. Such participation also represents an opportunity to demonstrate both interest in and competence to contribute to the organization's evaluation needs.

Over time, the best way to gain early awareness of evaluation RFPs consistently is to develop and make known a track record of outstanding evaluation work. As in any other walk of life, nothing succeeds like success. Frequently, RFP issuers will find evaluators who are known for their extensive and consistently high-quality evaluation services rather than vice versa.

Considerations in Deciding Whether to Respond to an RFP

We also note from our experience that if one learns about an evaluation RFP only after it has been published, it may be a waste of time and effort to write a proposal. In such situations, the time to respond may be very short. Also, other respondents may have been privy to the RFP during its development and have a long head start in developing their proposal. For many of the evaluation contracts we have won, we have been in the latter position. Although such situations do not constitute a level playing field, they are a part of the real world of RFP competitions. Nevertheless, many evaluation RFPs attract no or only a few evaluation proposals. Thus, it is not always a bad idea for an evaluator to bid on evaluation RFPs that he or she only learned about in a publication.

We also note that there have been RFP cases where the subject evaluation was "wired" to a particular respondent. Thus, many respondents to a fictitious RFP wasted their time in writing

an evaluation proposal because the preferred evaluator always had odds stacked in his or her favor. Possibly this evaluator had excelled in conducting previous evaluations of the subject program, had established a valuable database on the program, had acquired and maintained a staff with just the right qualifications to proceed with subsequent evaluations, and had earned the confidence of the program sponsor and other stakeholders. Understandably, the client organization wanted to sustain and build on its past investment in this evaluation contractor. Although the organization was compelled by statutory or other reasons to seek bids, it was always predictable that it would hire the evaluator of record again. The sponsor might well have written the RFP so that the evaluator of record would be the obvious choice—for example, because that evaluator exceeded any other party's experience in evaluating the subject program and had staff with just the right combination of qualifications. We cite this type of dubious practice as a part of the real world of evaluation RFPs. We advise evaluators to be alert to such situations so as not to waste time bidding on wired evaluation "opportunities."

Questions to Ask in Assessing an RFP

In looking at any evaluation RFP, potential bidders should carefully scrutinize the opportunity. Questions to ask include the following:

- Does this program's evaluation history reveal that the sponsor has had a sustained, successful relationship with a particular evaluator who is likely or at least eligible to bid on this evaluation?
- Is the timeline for responding unusually short?
- Does the RFP spell out criteria for selecting a bidder that almost excludes any party other than the evaluator of record?
- Is the content of the RFP built largely on evaluation plans and reports that were authored by the evaluator of record?
- Does the RFP essentially require the precise methodology that the evaluator of record employed in previous evaluations of this program?

If the answers to these questions are all or mainly yes, then it might be prudent to forgo responding to the particular evaluation RFP.

Evaluation RFQs

An evaluation RFQ is similar to an RFP in all matters except for the level of openness to different possible evaluation designs. Whereas an evaluation RFP asks the respondent to propose a plan for conducting the subject evaluation, an RFQ usually stipulates the methodology to be employed and asks the respondent either to quote a price for conducting the specified study or to submit his or her qualifications to conduct the study. The prescribed elements of design in an RFQ almost always are highly specific and leave the successful bidder little room for creativity and discretion.

From our vantage point, RFQs usually are not attractive options for applying one's evaluation skills. They place the evaluator essentially in a technical role and may prevent explorations that are necessary to assess a program's merit and worth. We acknowledge that the given evaluation design might have been developed carefully and appropriately; that the sponsor appropriately may seek out an evaluator to faithfully execute the design; and that providing such an evaluation service is legitimate, even if not creative.

Questions to Ask in Assessing an RFQ

Before pursuing an RFQ, an evaluator should address such questions as the following:

- Is the prescribed methodology appropriately responsive to the full range of important questions concerning the program's merit?
- Will competent implementation of the stipulated methods assuredly expose a failed program as well as hail one that succeeded?
- Does the prescribed methodology include an appropriate range of qualitative as well as quantitative methods?
- Is the prescribed methodology unbiased in looking at both strengths and weaknesses?
- Will implementation of the stipulated reporting plan ensure that findings are accessible to all right-to-know audiences?
- Does the prescribed methodology allow access to all relevant sources of information about the program?
- Will the prescribed methodology allow the evaluator to conduct an evaluation in which he or she can take pride?
- Will implementation of the prescribed methodology meet the standards of the evaluation field?

To the extent that an evaluator has to answer these questions in the negative, he or she might want to pursue other, better opportunities for applying his or her evaluation skills.

Internal Evaluation Assignments

Many evaluators are not independent contractors, but rather internal evaluators. They work within their organization and address its evaluation needs. Often those needs entail conducting evaluations of the organization's externally funded projects. Other times, internal evaluators assess certain programs or divisions within their organization.

Internal evaluations are vital to an organization's health and accountability. They are especially important for guiding program planning and improvement. In addition to requesting internal evaluations, the organization often has to bring in outside evaluators, who are more independent than the insiders. However, it is internal evaluators who by and large provide the information that the outside evaluators use to reach their conclusions and judgments.

Thus, internal evaluators have an important role in helping the organization maintain its accountability, even when outsiders conduct and report on the evaluations.

Advice for Setting Organizational Evaluation Priorities

In an organization of any size, there are more evaluation needs than the internal evaluators can address. Therefore, it is important that the organization have a process for assigning evaluation priorities, allocating evaluation resources, and scheduling the work. We think the internal evaluation team should annually assess the organization's evaluation needs, work with the organization's hierarchy to set evaluation priorities, and develop and carry out an annual program of internal evaluations. In addition to planning and scheduling internal evaluations, it is also important for an organization to maintain an evaluation contingency fund by which to address those emergent and important evaluation needs that were not predicted.

One way to set priorities and also foster use of evaluation findings is to establish a stakeholder evaluation review panel. The members should be representative of the organization's structure both horizontally and vertically. Such a panel's responsibilities could be to review annual assessments of evaluation needs, help set annual priorities for allocating evaluation resources, review evaluation plans and reports, help promote use of evaluation findings, and help develop the organization's evaluation policies and procedures.

Clearly, internal evaluators face a difficult obstacle in the form of their natural conflicts of interest. As professional inquirers, they need to issue valid assessments of merit and worth. Yet they also have to contribute to their organization's welfare and not cause it to fail or experience undue embarrassment, as might be the case in issuing and disseminating negative reports. For internal evaluators to walk the fine line between valid, forthright evaluation and advocacy for their organization's welfare, we think the organization should adopt and follow the standards of the evaluation field. This implies that all decision makers within the organization, at all levels, must become as conversant with adopted standards as the evaluators. The internal evaluators should faithfully follow these standards in conducting their evaluations. If it is clear that they cannot do so in a particular evaluation case, then they should use the standards to convince their organization to contract with an outside evaluator or at least an outside metaevaluator. If it is not feasible to bring in an outside party, the internal evaluators should make clear in their report the problems they faced; how they addressed these; and what they see as limitations of their findings, together with the reasons.

Sole-Source Requests for Evaluation

Experienced evaluators often are pursued on a sole-source/noncompetitive basis to conduct evaluations. This can be a fortunate situation for such an evaluator for a number of reasons. First, there is a good prospect that the evaluation findings will be used, because the sponsor wants the evaluation done. Second, the evaluator usually will be given discretion in matters of evaluation design. Third, on rare occasions, the sponsor may inform the evaluator that he or she has, within reasonable limits, a blank check to fund all appropriate evaluation tasks. Fourth, the evaluator often will be allowed to set a reasonable timeline to accomplish the needed

work. Clearly, an evaluator who is pursued by a sponsor and given an exclusive evaluation opportunity will want to seriously consider the opportunity before turning it aside.

Nevertheless, there can be good reasons to reject such opportunities. Possibly the sponsor is seeking a good report—not in terms of quality, but in terms of a positive judgment of the evaluand—and is willing to pay a high price for it. Conversely, the sponsor may be seeking and willing to pay handsomely for an unmitigated indictment of a program. Or, more subtly, the sponsor may open the way for a professionally sound evaluation but plan to use any indication of a program's weakness to fire the director or cancel the program. We have seen examples of each of these in our evaluation work. As we argued in Chapter 5, evaluators should not be in the business of conducting pseudoevaluations.

Accordingly, we advise evaluators who are sought out by a sponsor to undertake a background investigation before signing on to do the evaluation. It is especially important to identify and have an exchange with persons and groups that might experience harm as a consequence of the evaluation. Often they can offer insights into any hidden agenda for the evaluation. In general, it is important to learn as much as possible about the political climate surrounding the evaluation request before signing on. If red flags appear, the prospective evaluator can decide not to proceed or to proceed only under contractual terms that protect the evaluation's integrity and safeguard the legitimate interests of program stakeholders.

An evaluator can increase the prospects of being sought out for evaluation assignments in a number of ways. Most important is to develop a track record of conducting technically competent and useful evaluations. It is also a good idea to publish lessons learned and conduct training sessions based on one's evaluations. Evaluators are wise to be of service to potential client organizations, for example, by helping evaluate evaluation proposals and develop evaluation RFPs. An evaluator can prepare and disseminate a brochure describing his or her qualifications, experience, and availability for evaluation work. In addition, the evaluator might study the annual reports of prospective evaluation clients and send them a letter indicating his or her availability and interest in evaluating their programs. It is prudent for the evaluator to schedule visits to prospective funding organizations and provide them with information that has relevance to their evaluation needs. Also, an evaluation organization can maintain a Web site that includes evaluation exemplars and information about the organization and its staff (for an excellent example, visit www.wmich.edu/evalctr/). To the degree that qualified evaluators or evaluation organizations conduct activities such as those identified, they are likely to have many evaluation opportunities essentially walk through their doorway. However, we emphasize again that before signing on, the prospective evaluator should carefully scrutinize any sole-source opportunity for the possibility of an inappropriate, hidden agenda.

Evaluator-Initiated Evaluation Opportunities

Experienced evaluators often develop a track record of conducting evaluations within a given domain, such as charter schools, computer technology, community development, best business practice, employment, science education, or digital imaging technology. As they proceed from evaluation to evaluation, they will see often see a need for an important study that could help

advance the area or perhaps help turn it in a new direction. Accordingly, such evaluators probably will not wait for a relevant RFP or other evaluation opportunity to emerge. Instead, they will act proactively to help generate an appropriate evaluation opportunity. The following scenario, based on many actual evaluator-initiated evaluations of which we are aware, illustrates how the proactive evaluator might proceed.

An evaluator might schedule a visit to an organization, such as a foundation or government agency, that has funded evaluations in his or her particular area of interest and that may have discretionary funds for evaluator-initiated evaluations. In the course of scheduling the visit, the prospective evaluator might send a brief letter noting the need for evaluation in the substantive field and his or her desire to explore that need and how it could be addressed. (The initial contact could also be an informal encounter between the evaluator and a representative of the funding organization at a professional meeting.) The evaluator might take along to the scheduled meeting some brief written material, such as a list of talking points. However, he or she would be smart not to present, during this initial meeting, anything like a full-blown evaluation plan. Instead, it is better to establish rapport with the funding organization's staff and engage in a give-and-take exchange about the need for evaluation and how best to address it. At the meeting's conclusion, the evaluator probably would suggest and secure agreement on appropriate next steps. Typically the evaluator would send the funding organization a summary of the initial meeting, including any consensus that was reached, plus a draft evaluation plan. Subsequently the evaluator might engage in one or more follow-up meetings so that he or she and personnel of the funding organization could go over and strengthen the evaluation plan. If all goes well in the process of exchange and collaborative planning, the funding organization may ask the evaluator to submit a formal proposal for sole-source funding.

In summary, the evaluator should pursue a process of interaction and development of mutual understanding prior to detailing the evaluation plan, a process that could require months. Throughout the process of dialogue and deliberation, the evaluator should document the exchanges, and after each meeting should send a record of what was discussed, including any key agreements on next steps.

Cultivation of funding organizations followed by collaborative development of evaluation plans is a close-to-ideal way for evaluators to pursue a line of evaluation work. Of course, the evaluator must protect the integrity of the evaluation plan and process and not afford the sponsor inappropriate control over the evaluation procedures or reports. The evaluator should ground the evaluation in standards of the evaluation field, attest at appropriate points in the evaluation process to the extent to which the evaluation is meeting the standards, and do everything possible to ensure that the evaluation is subjected to an independent metaevaluation. By pursuing such safeguards, the evaluator should be able to retain an appropriate level of independence in the evaluation work while enjoying the benefits of a functional working relationship with the funding organization. If at all possible, the evaluator should obtain a grant rather than a contract or a cooperative agreement. If appropriate safeguards have been instituted, however, a defensible evaluator-initiated evaluation can be conducted under any of these arrangements.

Bidders' Conferences

In the case of a relatively high-cost evaluation, the sponsor often announces and conducts a bidders' conference. The conference's purpose is to give all potential bidders an equal opportunity to receive background information about the needed evaluation and address questions to the sponsor's representatives. The conference typically is conducted in an auditorium and runs for one to two hours. Conference leaders will be closely scripted in terms of the questions they can and cannot answer. The conference begins with an overview of the needed evaluation and the bidding requirements. Often the presenters distribute materials to supplement the evaluation RFP. The bulk of the meeting follows a question-and-answer format. Usually this segment is tape-recorded, with a transcript sent to all those in attendance and others if they request it. In addition, the sponsor will distribute a list of all conference attendees.

Although attendance at the conference is not a condition for entering or winning the RFP competition, there are several advantages to attending. First, an evaluator can make sure that his or her most important questions are asked. Second, it is always of interest to see who is in attendance. Observing who asks which questions and how different attendees interact before and after the session may help the evaluator size up the competition. Also, by attending the conference and interacting with some of the participants, an evaluator can consider possible advantages of partnering with other attendees to make a collaborative proposal.

In attending the conference, it is important to remember that the attendees are potential competitors. Before, during, and after the conference, they are likely to seek information from or about other evaluators and their respective organizations that could help them win the proposal competition. Thus, attendees must be wary of disclosing proprietary information that would help a competitor. Such information could include whether one has decided to bid, who is likely to lead the bidding effort, what other staff members will be involved, what consultants are being sought, what one's history is in regard to the particular RFP, what one considers to be a probable dollar cost for the evaluation, whether one would collaborate with another organization, what background planning has already been done, and what political support has been lined up. It is prudent never to volunteer information to others on any such matters and not to ask questions in the public meeting that would reveal information that could advantage the competitors. During the meeting, an evaluator usually is wise to wait before posing questions to see if other attendees ask those questions. This is because posing questions can expose one's plans for responding to the RFP. One should, however, listen intently during the meeting and before and after it for information helpful in writing a winning proposal. Also, it is important to take good notes based on attendance at the meeting, because the transcripts might not be complete and forthcoming in a timely fashion.

The preceding discussion of gamesmanship in responding to evaluation RFPs may seem distasteful. In fact, we found it distasteful to have to write about it. That being said, evaluations occur in a political context. It would be naive not to consider and effectively address the political realities of competitions for evaluation projects. To make this error would consistently put one on the losing side. It would also be wasteful of the invested time and funds.

Summary

In this chapter we have offered leads about how best to find, assess, and address evaluation opportunities. Evaluators may uncover and pursue a wide range of evaluation opportunities. These include RFPs, RFQs, internal evaluation assignments, sole-source evaluations, and evaluator-initiated opportunities. We have stressed that evaluators, in responding to or generating such opportunities, should carefully assess whether potential opportunities are worth pursuing in terms of both feasibility and ethicality. We have emphasized that evaluators should always hold their evaluations to the standards of the evaluation field and that they should seek to have their evaluations subjected to independent metaevaluations. We have also given our perspective on how best to participate in bidders' conferences.

REVIEW QUESTIONS

1. What is an evaluation RFP, and what are the sources of RFPs?
2. What are signs that you would have a poor shot at winning an evaluation RFP competition?
3. What is meant by the observation that an evaluation RFP is "wired," and what are the signs that this is the case?
4. Why is it important to consider whether an RFP calls for a grant, a contract, or a cooperative agreement, and why would an evaluator usually prefer a grant?
5. What are the hazards of entering into a cooperative agreement, and what steps can an evaluator take to protect the integrity of an evaluation under these circumstances?
6. What is an RFQ, and why might it not appeal to an evaluator's creativity?
7. What are evaluation review panels, what should be the membership of such panels, and what is their role in internal evaluation systems?
8. What is a sole-source request for an evaluation, and what are possible reasons to reject such an opportunity?
9. What is an evaluator-initiated evaluation opportunity, and what steps could an evaluator follow to effectively generate such an opportunity?
10. What is a bidders' conference, what are the advantages of attending one, and what are some cautions associated with one's behavior at the conference?

Group Exercises

Exercise 1

Outline a strategy that a neophyte evaluator could follow to consistently learn about evaluation opportunities.

Exercise 2

Suppose that you are outlining a policy to address conflict-of-interest issues in an organization's internal evaluation system. Define these potential issues, and list safeguards the organization could institute to address them effectively.

Exercise 3

What are the advantages of responding to a sole-source request for an evaluation, what are potential threats to such an evaluation's integrity, and what can the evaluator do to protect the evaluation's integrity?

Exercise 4

What are some of the basic precautions an evaluator should observe when considering responding to any evaluation RFP?

Suggested Supplemental Reading

Hamper, R. J., & Baugh, L. S. (2011). *Handbook for writing proposals* (2nd ed.). Columbus, OH: McGraw-Hill.

CHAPTER 18

FIRST STEPS IN ADDRESSING EVALUATION OPPORTUNITIES

After deciding to pursue a program evaluation opportunity, one must engage in an array of start-up activities: defining evaluation staffing needs, recruiting team members and collaborators, defining staff evaluation assignments and arranging to give credit to staff members for their contributions, developing thorough familiarity with the need for the evaluation, stipulating the standards for guiding and assessing the evaluation, establishing an institutional base of support for the projected work, satisfying institutional requirements for protecting the rights of human subjects, obtaining such appendix materials as letters of support and institutional as well as individual vitae, and planning for a stakeholder evaluation review panel. These activities are preliminary to the detailed work in developing the technical evaluation design, creating an appropriate budget, drafting a contract to cover the evaluation work, and packaging and submitting the evaluation proposal materials.

Often the prospective evaluator will need to pursue the initial start-up tasks expeditiously, especially in responding to a request for competing evaluation proposals (that is, a request for proposal [RFP]). Three main reasons underlie the need to move ahead proactively and promptly. First, the evaluator needs to recruit the most qualified staff, consultants, and (as appropriate) collaborating organizations before the competition lines them up. Second, he or she needs to draft evaluation proposal materials early, so that successive drafts can be prepared and critically reviewed, and so that ultimately a highly competitive final proposal is prepared ahead of the deadline for submission.

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **Determining and defining roles for a needed evaluation team**
- **Recruiting evaluation team members**
- **Subcontracting part of the evaluation work**
- **Clarifying the need for the evaluation**
- **Meeting the evaluation field's standards**
- **Arranging for institutional support**
- **Giving credit to evaluation team members for their contributions**
- **Satisfying requirements of a human subjects institutional review board**
- **Developing an evaluation proposal's appendix**
- **Setting up a stakeholder evaluation review panel**

Third, he or she needs to provide ample time to channel the proposal materials through a human subjects institutional review board and to acquire all the needed support and signatures from his or her institution.

In this chapter, we define some of the initial start-up activities and offer our advice. In ensuing chapters, we address other evaluation start-up activities, particularly evaluation design, evaluation budgeting, and evaluation contracting. We recognize that this chapter (and the other Part Four chapters) is especially applicable to relatively large-scale evaluations that require a team of participants. Nonetheless, many of the lessons apply to small studies conducted by a single evaluator.

Developing the Evaluation Team

One of the highest-priority start-up activities is to begin determining prospective evaluation participants and obtaining commitments from them. Having decided to proceed with the evaluation, the initiator should have a firm idea of the needed evaluation expertise. Example roles for a particular evaluation could be evaluation designer and manager, subject matter specialist, field data collector, measurement and analysis specialist, communication specialist, editor, and secretary. Full-time staff members might fill certain roles, and part-time consultants could carry out other roles. Often, certain evaluation team members will carry out more than one role. A large-scale evaluation might also require collaboration with one or more other organizations.

Recruitment of Evaluation Participants

After identifying the needed evaluation roles, the evaluator should proceed with all due haste to recruit members for the evaluation team. Experienced evaluators often have in mind a range of highly qualified persons with the needed expertise. Even experienced evaluators, however, often contact trusted colleagues for recommendations and look into the literature for persons who have published in the relevant substantive and technical areas. These are useful moves.

After listing potential evaluation participants, it is time to start contacting them in person, by telephone, or less desirably by e-mail. In the ensuing exchanges with each possible participant, the initiator should identify the evaluand, the sponsor, the evaluation's purpose, the main evaluation questions, the projected evaluation approach, the timeline, the role and amount of time envisioned for the person (or organization), and the expected level of compensation for the needed service once the evaluation is funded. Moreover, the evaluation's context, and in particular any abnormal political climate, should be explained and discussed. The initiator should respond to the potential recruits' questions and be open to hearing and using their ideas. For recruits who are willing to commit, the initiator should request a copy of their résumé and a letter stating their willingness to participate once the evaluation is funded. These materials will be placed in the proposal's appendix. Also, the initiator should keep all the recruits informed as the evaluation planning proceeds.

Developing Thorough Familiarity with the Need for the Evaluation

Prior to deciding to pursue the evaluation opportunity, the initiator will have learned a good deal about the evaluand and the need for the evaluation by studying the RFP (if there is one) and possibly attending a bidders' conference. But such activities are only the beginning of what has to be done. In addition, the initiator should search out relevant materials and people who know a good deal about the situation. Relevant materials might include past evaluations of the evaluand, journal articles, newspaper clippings, and a Web site. Key informants could include persons who previously evaluated the evaluand, experts in the subject matter area, and persons who have conducted and published research in the program area. As relevant materials are identified, the initiator and collaborators should study and discuss them and place them in an evaluation project library. Similarly, they should hold discussions with key informants, including those who may be expected to oppose the program or the evaluation. In general, the initiator and colleagues should learn all they can relevant to the evaluation assignment.

Stipulating Standards for Guiding and Assessing the Evaluation

In Chapter 3 we presented three sets of evaluation standards: the Joint Committee on Standards for Educational Evaluation's *Program Evaluation Standards,* the American Evaluation Association's *Guiding Principles for Evaluators,* and the U.S. Government Accountability Office's *Government Auditing Standards*. We also explained the fundamental importance of standards for guiding and assessing evaluation work and helping confirm its credibility. Depending on the particular evaluation situation and the initiator's preference, we think any of these three sets of standards can provide an appropriate foundation for a program evaluation. We advise the initiator who is planning an evaluation to select, in communication with the client, one or more of these sets of standards and present them to the evaluation team as the guiding policy for the contemplated evaluation. The initiator should require all evaluation team members to learn and apply the selected standards. These standards should be made an explicit part of the evaluation proposal.

Establishing Institutional Support for the Projected Evaluation

An early task for an evaluation initiator in preparing to submit a proposal is to gain support from his or her organization. The initiator should inform relevant superiors, administrative officials, and colleagues of the plan to write and submit a proposal; the timeline for completing and delivering the proposal; and the needed institutional resources, sign-offs, and assistance. At this early stage, administrators at the initiator's institution should be informed of the amount of money believed to be available for the evaluation work. The initiator should also be frank

in discussing the feasibility of submitting a winning proposal, given the institution's likely requirements for indirect/institutional overhead cost reimbursement. In some cases, it may be feasible for the institution to provide some type of matching support, such as a reduction in the indirect cost rate or contributed time of one or more evaluation staff members. Discussions of institutional support should also include such matters as needed release time for certain of the institution's staff members and possible subcontracts with other organizations whose services will be needed. In regard to subcontracting, it will be important to talk early with the lead organization's attorney who would be involved in writing and approving any needed subcontracts.

Arrangements for Giving Staff Members Credit for Their Evaluation Contributions

A point often overlooked is the need for the parent organization to commit to providing staff members who participate in the evaluation with due credit for their contributions to the evaluation. Staff members' excellent performance on a contracted evaluation should count toward their salary increases, promotions, awards, and, as relevant, tenure. It is wise to work out in advance how the organization will provide evaluation participants with just recognition and rewards, such as added salary.

Engaging the Human Subjects Institutional Review Board

Many organizations have a human subjects institutional review board (HSIRB) that typically has the authority to prevent a proposal from going forward if it does not satisfy the board's standards. Such boards have forms to fill out and often have a time-consuming review process that can be onerous. Clearly, the initiator of an evaluation proposal should contact the review board early and make sure the members understand his or her intention to submit a proposal plus the time frame. The initiator should fill out and submit review board forms as soon as possible. More than one excellent evaluation proposal has been rejected or has not been pursued because it failed to meet review board requirements in a timely fashion. The evaluator can make a strong case to the HSIRB that the evaluation will meet pertinent ethical requirements for observing and protecting the rights of human subjects by documenting that the evaluation will be designed and conducted to meet the evaluation field's standards.

Developing the Evaluation Proposal's Appendix

Unfortunately, many evaluation planners wait until the last moment to compile an evaluation proposal's appendix of essential background information. Consequently, some needed materials may not be included, and the included materials may be superficial or poorly prepared. Such weaknesses in an evaluation proposal can substantially worsen its prospects for funding.

From the start of preparing to submit a proposal, the initiator should begin soliciting and compiling the appendix materials. These may include, among others, a summary of the adopted

evaluation standards, an institutional vita, personal résumés or curricula vitae for personnel, a list of members of an evaluation review panel, and letters of commitment.

It is important to scrutinize and be selective in regard to what goes in the appendix. For key staff members, full-length résumés may be important, whereas summaries may suffice for less crucial participants. In obtaining letters of commitment, it can be useful to provide the letter writers with a model letter of commitment. This needs to be done as early as possible, so that the initiator can follow up to obtain letters from late respondents. In further developing the evaluation design, it may also be possible and important to include sample evaluation instruments.

Planning for a Stakeholder Review Panel

In many evaluations, it can be important to arrange for the involvement of a stakeholder evaluation review panel. This panel's tasks should include reviewing and critiquing draft evaluation materials; helping to disseminate evaluation findings; and, as appropriate, facilitating data collection. The panel's membership could include staff members of the program being evaluated, constituents of the program, relevant policymakers, evaluation experts, and persons from the organization that funds the subject program. Effective employment of such a review panel can help ensure that evaluation instruments are understandable to data providers, that reports are responsive to key questions and clear, that approaches to data collection are feasible and efficient, and that evaluation findings are heeded and used by the intended audience.

Based on our experience, we believe it is important not to label this panel as an "advisory panel." The type of panel we have in mind is capable of reviewing draft evaluation plans, reports, and tools for clarity, relevance, political sensitivity, and utility. Many of its members, however, would not be qualified to judge the technical merit of such methodological matters as evaluation design, sampling, data collection, and statistical analysis. In any case, it would be a mistake for the evaluator to give the impression that the review panel has authority for deciding how the evaluation will be conducted. We think evaluators should make clear to members of the review panel at the outset that their task is not to help design the evaluation's technical aspects, but to provide periodic reviews of how well and how diplomatically draft reports are being communicated to the intended audience and to comment on how data collection can best be scheduled and carried out in the program's environment. The evaluator must retain the authority over as well as the responsibility for using stakeholder feedback to decide how best to proceed with an evaluation and how appropriately to manage and refine along the way the evaluation's implementation.

Summary

In this chapter we have noted some of the key early steps in developing a winning evaluation proposal. The crucial start-up steps include recruiting and defining roles for evaluation team members and possible collaborators, clarifying the need for the evaluation, adopting standards for guiding and assessing the evaluation, arranging institutional support for the evaluation,

arranging to give credit to evaluation team members for their contributions, meeting requirements of a human subjects institutional review board, developing the evaluation proposal's appendix, and setting up a stakeholder evaluation review panel. These steps are essential in establishing a strong foundation for the projected evaluation. Evaluation planners who give short shrift to such start-up tasks diminish their prospects of submitting a winning evaluation proposal.

REVIEW QUESTIONS

1. In general terms, what are appropriate labels for and brief definitions of five essential start-up activities for large-scale program evaluation?
2. What key staff roles have to be defined and carried out in a relatively complex program evaluation?
   a. What brief definition would you assign to each role?
   b. Is it necessary that a different person carry out each role?
   c. Why or why not?
3. What is this chapter's stance on giving credit to evaluation team members for their contributions to an evaluation?
   a. What is your assessment of the pros and cons of this stance?
   b. What is your assessment of the relative feasibility of realizing this stance in organizations with which you have experience (for example, universities, school districts, government organizations, private foundations, or private companies)?
   c. Briefly explain your assessment in regard to each of the organizations you listed for the previous question.
4. What are at least three concrete steps for clarifying the need for an evaluation?
5. In reference to the program evaluation standards listed in Chapter 3, which particular standards would you judge to have relevance for protecting the rights of human subjects?
6. What are at least four types of institutional support that are relevant to preparing an evaluation proposal for external funding?
7. Many organizations have a human subjects institutional review board.
   a. What is the role of such a board?
   b. What is the likely consequence of bypassing the board in the process of submitting an evaluation proposal for external funding?
   c. Do you think such a board is expected to assess and approve a proposed evaluation's methods?
   d. If yes, why? If no, why not?
   e. Why would an evaluation's incorporation of the evaluation field's standards help the evaluation proposal meet the requirements of a human subjects institutional review board?

8. What are at least four kinds of relevant materials pertaining to a program that an evaluation initiator should obtain and study before finalizing an evaluation proposal?
9. What are at least three reasons why all participants in an evaluation should be conversant with the standards that will be used to guide and assess the evaluation?
10. When should an evaluator compile an appendix of essential background information for an evaluation proposal? What general items would you include in a tentative outline for an evaluation's appendix?

Group Exercise

A charitable foundation has issued an RFP for a longitudinal, approximately five-year evaluation of its self-help housing program. The program's main features are outlined as follows:

- Located in a poverty-stricken area of a large city
- Construction to be on a seven-acre plot recently donated to the foundation
- Helping thirty-two low-income families obtain low-cost, thirty-year home mortgages
- Selecting families with at least one parent who is gainfully employed, clean criminal background checks, an acceptable credit rating, young children (twelve and under), and a clear need for decent housing
- Houses with three bedrooms, fifteen hundred square feet, an attached garage, and a small yard
- Guiding each family through the process of constructing its new home
- Families obtaining their own hand tools
- Foundation subsidy by providing each family with a lot; the needed concrete, electrical, roofing, and plumbing work for the houses; needed infrastructure, including streets, sewers, water, electrical service, and garbage pickup; an on-site coordinator of builders and cobuilders, and two on-site construction experts to provide on-the-job construction training and support; periodic foundation-sponsored social events to promote cohesiveness among the families; plus periodic special self-improvement courses and counseling for both adults and children
- Covenants, requiring no farm animals plus freedom from domestic violence and drug and alcohol abuse
- Each builder and a cobuilder (usually the spouse) required to work on the houses for ten hours during each Saturday and Sunday, over a twelve-month period
- Four increments of eight houses to be built successively during a total development period of four years

- Involved families responsible for obtaining child care while working on their respective houses

The RFP requests both ongoing formative evaluation and a five-year summative evaluation. The specified formative evaluation questions are as follows:

1. How well is each of the program's main features being carried out?
2. What problems are being encountered in the Saturday and Sunday construction processes?
3. What are the families' common and idiosyncratic needs in regard to housing?
4. Is this program meeting those needs effectively?
5. Are relations among the builders harmonious during the Saturday and Sunday construction processes?
6. Over the course of the program, are children's needs being met, or is the program proving counterproductive for any of them?
7. Along the way, what improvements in either program design or implementation are needed?

The specified summative evaluation questions are as follows:

1. To what extent did the program succeed in meeting the housing needs of the involved thirty-two families?
2. To what extent did the program succeed in meeting pertinent needs of the participating children?
3. To what extent did each of the program's planned features prove sound and effective?
4. What were the program's effects on the participating families—positive and negative, planned and unplanned?
5. What were the program's effects on the surrounding community?
6. What were the program's most important strengths and its most important weaknesses?
7. Overall, what was the program's level of cost-effectiveness?

As members of a university's evaluation center, you and several colleagues have decided to develop and submit a proposal for evaluating the foundation's self-help housing program. In considering how to address this evaluation opportunity, respond to the following questions:

1. What do you see as the appropriate composition and particular roles of the needed evaluation team?
2. What additional information would you seek to clarify the need for this evaluation, and where do you think your group might find this information?
3. What set of standards would you choose to undergird the evaluation? Considering those standards, would you add any evaluation questions to the lists provided by the foundation?
4. What types of support would you request from your university to submit a viable, winning proposal, and what is your rationale for each request?

5. What specifically would you ask the university to do to give credit to the evaluation team members for their contributions to the evaluation, especially the involved faculty members?
6. What methods would you employ to conduct the formative evaluation?
7. What methods would you employ to conduct the summative evaluation?
8. In general, what argument would you offer the university's HSIRB to demonstrate that your evaluation will protect the rights of human subjects?
9. What main items would you include in the evaluation proposal's appendix?
10. Would you set up a stakeholder evaluation review panel? Why or why not? If yes, what roles would you include on the review panel?

Suggested Supplemental Readings

Bhola, H. S. (1998). Program evaluation for program renewal: A study of the National Literacy Program in Namibia (NLPN). *Studies in Educational Evaluation*, *24*, 303–330.

Hamper, R. J., & Baugh, L. S. (2011). *Handbook for writing proposals* (2nd ed.). Columbus, OH: McGraw-Hill.

CHAPTER 19

DESIGNING EVALUATIONS

Having decided to conduct a study, the evaluator needs to prepare an appropriate design. An evaluation design is a set of decisions required to carry out the needed evaluation. These focus especially on determining the evaluand, identifying channels for informing and involving right-to-know audiences, defining questions to be addressed, clarifying relevant values and criteria to be applied, identifying information to be collected, specifying information collection and analysis tools and procedures, arranging data control protocols, planning for synthesizing findings, scheduling interim and final reports, determining reporting methods, taking steps to promote and support use of findings, and administering the evaluation. On a practical level, the evaluator needs to make design decisions before the evaluation work begins because these provide the basis for budgeting, contracting, staffing, and scheduling the needed work. In randomized experiments, the initial core design decisions are considered fixed because the evaluator seeks to hold treatment and control conditions separate and constant to identify their differential effects on assigned treatment and control groups of subjects. In the more general case of program evaluations, initial design decisions often must be reconsidered or fleshed out as the evaluation unfolds. This is especially so in formative and responsive evaluations. In such evaluations, the evaluator expects information needs to evolve as interim reports surface new issues, as the subject program matures (or falters), and as the client and other stakeholders raise new questions. Even in field experiments, contextual dynamics and needs and actions of experimental subjects may erode the evaluator's control over treatment and control conditions and cause the evaluator to modify the experimental design, or even replace

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **The definition of evaluation design**
- **Skills needed for preparing sound evaluation designs**
- **The details of an evaluation design based on the context, input, process, and product (CIPP) model**
- **An illustration of integrating evaluation standards into an evaluation design**
- **An illustration of inclusion of the advocate teams technique in an evaluation design**
- **The contents and uses of a generic checklist for designing evaluations**
- **The details of evaluation design components, including focusing the evaluation; collecting, organizing, and analyzing information; reporting interim and final results; and administering the evaluation**

it with a nonexperimental approach. In general, evaluators should periodically revisit, update, and delineate evaluation design decisions in consideration of evolving study conditions and client or other stakeholder needs. We advise readers to fix firmly in their minds that evaluation design is both process and product: initial design decisions appropriately are often general and tentative and become increasingly specific as an evaluation unfolds.

Over the course of an evaluation, the evaluator must exercise excellent communication and negotiation skills, responsiveness, and technical expertise in reaching and evolving sound design decisions. Overall, the evaluation design should address an audience's information needs, provide for judging the evaluand's merit and worth, be true to the evaluator's chosen evaluation model or approach, be capable of execution in the evaluand's setting, and in general meet the standards of the evaluation field. To address such challenges effectively, an evaluator requires an appropriate repertoire of qualitative and quantitative methods, competence in planning and administering evaluations, the ability to meet evaluation standards, political skills, and a good measure of creativity.

In addressing the topic of evaluation design, first we present and discuss a fictionalized example of an evaluation design that is based on an actual evaluation. We have focused on a fairly complex evaluation because it provides a basis for looking at a wide range of evaluation tasks and methods, in both this and subsequent chapters, and because it illustrates the frequent situation in which an evaluation design starts out as a general plan and takes on specificity after the study is funded and launched. The evaluation was keyed to professional standards for evaluations and had profound effects on the client's decisions, and in those respects, it was exemplary. A bonus is that the example is an evaluation of a military personnel evaluation system that has the structural characteristics of a program evaluation (assessment of an interrelated set of goal-directed activities) and the content of personnel evaluation.

Second, we present and discuss a generic checklist for use in making evaluation design decisions or checking the adequacy of a completed evaluation design. This checklist is applicable to any of the wide range of defensible evaluation models and approaches and may be used in conjunction with the design recommendations included with a chosen evaluation model or approach. It is a tool for use in constructing an initial design, fleshing it out as the evaluation proceeds, and checking the adequacy of a proposed design. The checklist is useful to both evaluators and clients. We also reference particular checkpoints to apprise readers of the additional design decisions that the evaluation team had to make during its evaluation of the military personnel evaluation system.

A Design Used for Evaluating the Performance Review System of a Military Organization

The evaluation design example reviewed in this chapter is an evaluation of the system used by the U.S. Marine Corps (USMC) for evaluating the job performance of its officers, staff noncommissioned officers, and sergeants. The USMC commandant was dissatisfied with his organization's performance review system (PRS); had ordered the organization's personnel

department to obtain an independent evaluation of that system; and required completion of the evaluation and subsequent reform of PRS by the end of his term as commandant, which would occur soon. An official of this organization invited a particular evaluator to lead the needed evaluation. The evaluator subsequently prepared and submitted the general evaluation design described in the following subsections. The USMC took two months to process the proposal and ultimately approved it, along with a fixed-price award of about $440,000, leaving six months to complete the work. The evaluator and his team fleshed out this design in the course of conducting the evaluation, as will be discussed in subsequent chapters.

Task Order for the Evaluation

This example began when the USMC commandant's representative provided the evaluator with a task order and offered a sole-source contract. The key tasks were to (1) assess the strengths and weaknesses of the existing PRS, (2) identify and assess alternative personnel evaluation systems, (3) help design a preferred system, and (4) develop a comprehensive plan for implementing the recommended system. The USMC required the contractor to complete all four tasks within eight months, which as already mentioned, due to a lengthy award process and a fixed deadline for the final report, became a six-month period. Deliverables were monthly progress reports, a scheduled interim report for each task, and a final report. The USMC also directed the contractor to brief the sponsoring committee and its study advisory committee on each task report and the final report. Sitting on the sponsoring committee were eleven general officers, two sergeant majors, five colonels, four majors, and two captains. The members of the study advisory committee were a brigadier general, two lieutenant colonels, three majors, two captains, one sergeant major, and two civilian employees of the PRS. Separate sessions were scheduled to brief each committee on each report; these were to be conducted at USMC headquarters in Washington, DC. Each printed report was to be delivered to the USMC a minimum of ten working days prior to the scheduled briefing sessions. Given the extensive amount of needed work, the short timeline, and the significance of the problem, the sponsor set no limit on the funds to be allocated to this evaluation project. The evaluator could request whatever amount of funding was required to do the job well and on time. In turn, the USMC would issue a fixed-price contract for the requested amount. The USMC would make available to the contractor all relevant information concerning its PRS, assign staff officers to support and act as liaisons for the evaluation project, provide meeting space and equipment at USMC headquarters, provide access to enlisted personnel and officers for interviews at the Quantico headquarters, and provide the needed funds.

Especially noteworthy in the task order was the appointment of the two USMC panels to read and react to all evaluation reports, the requirement that the lead evaluator and his team brief the panels on the reports, the explicit schedule of briefing sessions, and the fact that each panel was to be chaired by a high-ranking general officer. These highly responsible provisions by the study's sponsor did much to ensure that the USMC audience for the evaluation would critically review and use findings as appropriate. Although it was also helpful that the USMC allowed the evaluator and his team to interview marines, it was a decided limitation that this

had to occur at the Quantico headquarters, where the marines were under the close scrutiny of military leaders and might be expected to be less than candid about strengths and weaknesses of PRS procedures and leadership. As is evident in subsequent chapters, the evaluator later sought and secured approval to observe PRS operations and interview marines at other bases, which yielded more candid responses than those obtained at the USMC Quantico post. It is also noteworthy that the task order required the contractor to conduct evaluations of PRS and alternative personnel evaluation systems and to produce plans for responding to the evaluation findings. Although evaluators often prefer only to evaluate and not to recommend solutions, in this case combining the two types of tasks proved to be functional and in the interest of helping the USMC improve its personnel evaluation system.

Need for the Evaluation Project

The USMC had used the subject evaluation system for many years as the basis for retention, promotion, assignment, and mustering-out decisions. Congress periodically allocates finite numbers of positions at each rank in each military service. The distribution of positions, in each service, from lower to higher ranks approximates a pyramid for both enlisted and officer ranks. For example, in the USMC, there could be about nine thousand slots at the sergeant (E5) level but fewer than two hundred at the sergeant major (E9) level; analogously, second lieutenants (O1) could number about three thousand, compared with approximately four hundred colonels (O6). From the bottom to the top levels of ranks, there is increasing pressure to make room for new marines at each higher level. Of necessity, the USMC (and all other military services) employs an up-or-out promotion system. Because each higher-level rank has fewer slots than the immediately lower rank, not all meritorious marines can be promoted. This is especially so at higher ranks. After a certain number of years in his or her rank, theoretically each marine has to make room for a newcomer. (An exception is when enlistments in the service are not fully meeting the requirements of a certain military specialty. In such cases, a marine with the high-need specialty could be retained even if he or she failed a promotion review.) Depending on the needs of the service, after a defined period of time in a rank, a marine typically must be promoted or mustered out.

The USMC has a promotion board for each rank, and the boards must make the crucially important decisions about which marines to promote. They do so based largely on fitness reports prepared by each marine's immediate superior. The fitness report documents a superior's observations and assessment of the marine's performance, potential, and quality and is intended to be an accurate assessment of what is accomplished compared against job requirements. The supervisor is supposed to rate the marine based on missions, tasks, and standards that previously were communicated to the marine and also based on the marine's potential to serve at more senior levels and to accept ever-increasing responsibility. In making the ratings, the superior is expected to focus on known USMC values and the best of marine virtues and not on his or her personal preferences. At the lower ranks, almost all marines are promoted; for example, 95 percent of second lieutenants are promoted to first lieutenant, 88 percent of first lieutenants to captain, 75 percent of captains to major, 55 percent of majors

to lieutenant colonel, and 48 percent of lieutenant colonels to colonel. Such percentages change from year to year depending on the needs of the service and the availability of marines at each rank for consideration to be promoted.

Over the years, the supervising officers' ratings of marines had become highly suspect, and the PRS had fallen into disrepute. Criticisms of the system included unrealistic performance standards for promotion, a rating scale that yielded unreliable assessments, subjective ratings that were subject to bias, rampant inflation in ratings of performance, and the lack of a mechanism to audit ratings and correct invalid ratings. It had become common to rate marines who did not perform well down from outstanding to excellent. The lower levels of the scale were rarely, if ever, used. There was therefore little differentiation in the ratings. The pervasive grade inflation made it hard for the USMC promotion boards to discern which marines most merited promotion. The PRS had lost credibility with many marines, including some who had been promoted by the system to the level of a general officer. Marines throughout the USMC worried that promotion boards were using faulty evaluative information and making many poor or unjust promotions and mustering-out decisions. The suspected culprits were a faulty fitness report form and unreliable procedures for applying the form. Organization-wide concerns about the PRS were seen as impairing morale among the troops and possibly weakening the USMC's ability to fight and help win wars.

Evaluation Design

To investigate and address the problems in the PRS, the evaluator designed a project grounded in the Joint Committee on Standards for Educational Evaluation's *Personnel Evaluation Standards* (1988) and the CIPP evaluation model (see Chapters 7 and 13). The objectives that follow were identified as fully responsive to the evaluation needs underlying this project, but as extending beyond the constraints of the USMC's task order. Because the requested work was required to be completed within six months following contract approval, the project plan and budget realistically could be keyed only to completing the first four objectives and starting work on the fifth. The evaluator noted that based on the outcomes related to the funded evaluation project's first five objectives, as listed in the next subsection, he and his group would be willing to undertake a follow-up project to assist the USMC in fully achieving objective 5 and addressing objectives 6 through 14. He considered it important to apprise the USMC of the full scope of needed work, but not to promise more than could be accomplished within six months.

Objectives

Following is the full set of fourteen recommended objectives, with the first five providing the basis for the contracted evaluation work:

Foundation for the Project

1. Adapt and adopt the 1988 Joint Committee personnel evaluation standards as the official standards of quality for the PRS.

Context Evaluation

2. Evaluate the current PRS against the adapted personnel evaluation standards to identify strengths to be built on and problems to be solved or avoided.
3. Use the objective 2 results and relevant research and development literature to determine with appropriate USMC leadership the specifications for the new PRS.

Input Evaluation

4. Identify and develop alternative personnel evaluation systems and evaluate them against the personnel evaluation standards and against the specifications for the new PRS.
5. Assist USMC leaders to converge the best features of the alternative personnel evaluation systems into a sound design for the new PRS—including versions for evaluating the performance of officers, staff noncommissioned officers, and sergeants and provisions for auditing and correcting mistakes in individual personnel evaluations.

Process and Product Evaluations

6. Prepare a plan for testing and validating each version of the new PRS.
7. Train designated USMC personnel to field-test each version of the new PRS.
8. Conduct process and product evaluations to field-test the new PRS.
9. Evaluate the implementation and results of the field tests.
10. Assist appropriate USMC leaders in making needed corrections to the new PRS.

Institutionalization of the New PRS

11. Prepare the implementation resources for each version of the new PRS: manuals, instruments, report formats, funding plans, training materials, an appeals mechanism, and so on.
12. Design procedures for the transfer of current marines' performance assessment records into the new system.
13. Assist appropriate USMC leaders in setting up and installing an ongoing process for monitoring each version of the new PRS and improving it as needed.
14. Assist appropriate USMC leaders in setting up and installing an ongoing program for training personnel to implement each version of the new PRS.

Required Features of the New PRS

In general, it is wise to restate or characterize a potential sponsor's criteria for evaluating the evaluand, as stated in the task order or request for proposal (RFP). Such a recapitulation can reassure the potential sponsor that you are giving appropriate consideration to the criteria

the sponsor sees as important. Summarizing the sponsor's stated evaluative criteria also helps the proposal writer communicate about the possible evaluation in terms that the client will value and understand. Of course, the evaluator must not acquiesce to inappropriate criteria or necessarily limit the evaluation to only the sponsor's criteria. Given these provisos, the following is a characterization of USMC's criteria for a new PRS:

1. A clear framework for identifying marine duties to be assessed at several levels of experience and responsibility: supervisory, managerial, company and field-grade officer, and executive
2. A sound, workable procedure for articulating appropriate performance expectations for promotion and other personnel actions for each evaluee
3. Sound, workable procedures to ensure validity, reliability, objectivity, and creditability in appraising how well an individual meets performance expectations
4. Clear rules and procedures for identifying the authorized users and uses of appraisal results
5. Specifications to ensure that performance records are appropriate for the intended uses
6. Effective means for clear communication of the appraisal results to the evaluee and authorized users
7. Appropriate measures to make the transition into the new PRS and fairly consider records of personnel whose performance was reported using the current system
8. An effective mechanism to hear appeals, audit the ratings, and correct invalid findings
9. Safeguards against rating inflation
10. A mechanism for regularly assessing and improving each version of the new PRS

The evaluator saw these criteria for a new PRS as entirely appropriate, although not sufficient.

Standards of Sound Performance Evaluation

Moving beyond the sponsor-generated criteria, the evaluator recommended that the USMC adopt a comprehensive set of professional standards of sound performance evaluation for use in evaluating the existing PRS and alternative personnel evaluation systems, designing the new PRS, and periodically reviewing and improving the new PRS. The set recommended was adapted from the Joint Committee's *Personnel Evaluation Standards* (1988). The proposed standards require that performance evaluation systems be designed, implemented, and used to meet requirements of utility, propriety, feasibility, and accuracy. The specific standards recommended for each of these attributes are summarized in the text that follows. As seen here, the evaluation contractor provided parenthesized commentary to help explain some of the standards.

Utility

The recommended utility standards are intended to guide evaluations so that they will be informative, timely, and influential for use in strengthening personnel performance and making personnel decisions:

U1—Constructive Orientation. Performance evaluations should be constructive, so that they help the USMC develop human resources and encourage and help those evaluated to provide excellent service.

U2—Defined Uses. The users and the intended uses of a performance evaluation should be defined, so that the evaluation can address appropriate questions and supply the needed information.

(This standard requires the development of clear rules and procedures for identifying the authorized users and uses of appraisal results.)

U3—Evaluator Credibility. The performance evaluation system should be managed and executed by persons with the necessary qualifications, skills, training, and authority; and evaluators should conduct themselves in a professional, evenhanded manner, so that evaluation reports are respected and used.

U4—Functional Reporting. Reports should be clear, timely, accurate, and germane, so that they are of practical value to the evaluee, supervisor, and other appropriate users.

(This standard requires development of clear specifications to ensure that performance records are appropriate for the intended uses and that effective means are used to clearly communicate the appraisal results to the evaluee and authorized users.)

U5—Follow-Up and Impact. Performance evaluations should be followed up on, so that users and evaluees are aided to understand the results and take appropriate actions.

(This standard requires development and application of appropriate procedures for the transition to the new system and fair consideration of performance evaluation records of personnel whose performance was reported using the current system.)

Propriety

The recommended propriety standards require that evaluations be conducted legally, ethically, and with due regard for fairness to evaluees, users of evaluation results, and persons supervised and served by the evaluees:

P1—Service Orientation. Evaluations of marines should promote sound principles of democracy, fulfillment of the USMC's mission and objectives, and effective performance of duties, so that the USMC faithfully and effectively fulfills its constitutional obligations to the United States.

(According to this standard, performance evaluations should be planned, conducted, and used so that each marine is required and supported to effectively serve her or his country by carrying out assigned, appropriate duties, and so that, where indicated, sanctions that are in the best interest of the United States are enforced.)

P2—Formal Evaluation Guidelines. Guidelines for performance evaluations should be recorded in policy statements and performance evaluation manuals, so that evaluations are consistent, equitable, in accordance with pertinent laws and military codes, and effectively carried out.

P3—Conflict of Interest. Conflicts of interest should be identified and dealt with openly and honestly, so that they do not compromise the performance evaluation process and results.

(This standard reflects the fact that conflicts of interest are inherent in any system in which the supervisor evaluates the subordinate and must be controlled through effective mechanisms, such as the use of independent evaluators; complete, factual service records; self-reports; an appeals process; and regular monitoring and assessment of the evaluation system.)

P4—Access to Performance Evaluation Reports. Access to reports of performance evaluations should be limited to individuals with a legitimate need to review and use the reports, so that appropriate use of information is ensured.

(In accordance with this standard, there must be clear rules and procedures for limiting access to performance evaluation records to appropriately authorized persons.)

P5—Interaction with Evaluees. The evaluator should address evaluees in a professional, fair manner, so that their motivation, service reputation, self-esteem, and attitude toward performance appraisal are enhanced, or at least not needlessly and unfairly damaged.

(In keeping with this standard, performance evaluation should be conveyed and employed as a mechanism to enhance performance and pride in excellent service and to provide a fair basis for personnel decisions, not as a tool to intimidate, discourage, or mete out punishment.)

Feasibility

The recommended feasibility standards call for evaluation systems that are as easy to implement as possible, efficient in their use of time and resources, adequately funded to effectively maintain and improve evaluations, and viable within the program context:

F1—Practical Procedures. Performance evaluation procedures should be planned and conducted such that they produce needed information while minimizing disruption and cost.

(We interpret this standard as meaning that, wherever possible, the data collection activities for performance evaluation should be integrated into the ongoing process of supervision, maintaining personnel records, and personnel decision making.)

F2—Political Viability. Performance evaluation procedures should be planned and conducted such that representatives of all concerned parties are constructively involved in designing the system, testing it, and making it work.

(In keeping with this standard, it is important to keep interested parties informed about the professional nature of the development process through effective communication.)

F3—Fiscal Viability. Adequate time and resources should be provided for performance evaluation activities, so that evaluation plans can be effectively and efficiently implemented.

(To meet this standard, it is especially important to provide for the ongoing training, calibration, and monitoring of evaluators.)

Accuracy

The accuracy standards require that the obtained information be technically accurate, and that conclusions be linked logically to the data:

A1—Defined Role. The role, responsibilities, performance objectives, and needed qualifications of the evaluee should be clearly defined, so that the evaluator can gather valid assessment data.

(In accordance with this standard, role definitions should be derived from a clear, official framework for identifying marine duties at several levels of experience and responsibility: supervisory, managerial, company and field-grade officer, and executive. There also should be a sound, workable procedure for articulating appropriate performance expectations for promotion and other personnel actions for each evaluee. There should be procedures for reviewing and updating performance criteria as appropriate.)

A2—Work Environment. The context in which the evaluee works should be identified, described, and recorded, so that environmental influences and constraints on performance can be considered in the evaluation.

A3—Documentation of Procedures. The evaluation procedures followed should be documented, so that the evaluees and other users can assess the actual, in relation to the intended, procedures.

(In keeping with this standard, USMC leaders noted that it is especially important that evaluators be required to cite the duties evaluated, the evidence used to reach judgments and recommendations, how the evidence was obtained, and why that evidence is considered sufficient and credible.)

A4—Valid Measurement. The data collection and rating procedures should be chosen or developed and implemented on the basis of the described role and the intended use, so that the evaluator makes valid inferences about the evaluee's performance.

(To meet this standard, USMC leaders stipulated that these procedures should be a matter of record well before the performance evaluation is completed.)

A5—Reliable Measurement. Data collection and rating procedures should be chosen or developed to ensure reliability, so that the information obtained will provide consistent indications of the performance of the evaluee.

A6—Systematic Data Control. The information used in the evaluation should be kept secure and should be carefully processed and maintained, so as to ensure that the data maintained and analyzed are the same as the data collected.

A7—Bias Control. The evaluation process should provide safeguards against bias, so that the evaluee's performance is assessed fairly.

(Meeting this standard requires effective provisions to review evaluations, hear appeals, audit ratings, and correct invalid findings. USMC leaders stipulated that explicit safeguards against rating inflation must be built into the system.)

A8—Monitoring Evaluation Systems. The personnel evaluation system should be reviewed periodically and systematically against the preceding twenty standards, so that appropriate revisions can be made.

(Meeting this standard requires provisions for regularly assessing and improving the PRS and for explicitly protecting against grade inflation.)

At the outset of the evaluation, the USMC readily embraced the recommended standards and the associated parenthesized comments. It is noteworthy that it asked for only two changes in the standards, both in the area of feasibility. One change was to rename the Political Viability standard "Consensus Development." (USMC leaders did not want anyone to think they were trying to be politically correct.) The other change was to add a standard called "Transition to the New PRS." This added standard required provision for systematic adoption and installation of the new evaluation system. The evaluator and his team judged both changes to be sound and appropriate to the situation. The adapted standards then became the official USMC standards for assessing and improving the personnel evaluation system.

General Study Plan

The evaluator next presented a general plan for the evaluation. It stated that the evaluation project would be divided into five main tasks: (1) project organization and background analysis (weeks 1 through 6), (2) context evaluation of the current PRS (weeks 3 through 10), (3) input evaluation to identify and analyze alternative performance evaluation systems and literature review (weeks 7 though 18), (4) preparation and reporting of conclusions and recommendations for a new PRS (weeks 19 through 26), and (5) beginning efforts to plan for development and implementation of the proposed new PRS (weeks 27 through 34).

The evaluator projected that the evaluation project would provide the study sponsor with five main reports, corresponding to the tasks just listed, at a pace of about one per month. Evaluation team members were designated as E for the principal investigator and E1, E2, E3, E4, and E5 for other team members. The projected reports were as follows:

Finalized Project Plan

- Including preliminary background analysis of the PRS and proposed PRS standards
- Principal authors: E and E2
- For delivery during week 6

Evaluation of the Current PRS (Context Evaluation)

- Including a comparison of the current PRS to the PRS standards and proposed specifications for the new PRS
- Principal authors: E3 and E4
- For delivery during week 10

Evaluation of Alternative Personnel Evaluation Systems (Input Evaluation)

- Including descriptions of promising systems used in other branches of the military and in business and industry; a comparison of these systems to designated standards; and a literature review
- Principal authors: E1, E5, and E6
- For delivery during week 18

Conclusions and Recommendations

- Including conclusions about the reasons for the failure of the present PRS, the merits of alternative personnel evaluation systems, and a general design for the new PRS
- Principal authors: E, E1, E2, and E4
- For delivery during week 26

Plan for Development and Implementation of the Proposed New PRS (Including Process and Product Evaluation Plans)

- Including draft plans for operationalizing, field-testing, correcting, and installing the proposed new PRS
- Principal authors: E, E1, and E6
- For delivery during week 34

Project Personnel

The evaluation project tasks were to be performed by a central project team (the project director, project manager, and task group chairs) and three associated task groups (project management, context evaluation, and input evaluation), with the evaluator serving as project director. The central project team members, their project assignments, and their most pertinent areas of expertise were as follows:

E: Project director—personnel evaluation standards

E1: Project manager—project management

E2: Context evaluation task group chair—performance measurement

E3: Context evaluation task group member—statistics and computer technology

E4: Input evaluation/PRS alternatives task group chair—personnel psychology

E5: Input evaluation/PRS alternatives task group member—military personnel evaluation systems

The proposal included résumés for all proposed key project personnel. All members arguably were among the nation's top professionals in their respective specialties, and three of them had relevant military experience.

The central project team was designated to review and finalize reports from the three project task groups and ultimately to be responsible for reporting project conclusions and recommendations. This team was configured to include experts in personnel evaluation standards (E), personnel psychology (E4), performance measurement (E2), military personnel evaluation systems (E5), and project management (E1). The team's core responsibility was to serve as the project's working board. In addition to participating in team decision making, each member was given a major project task assignment.

The project management task group members were E, E1, a secretary, and two research associates. The evaluator (E) would oversee the work and ensure that it was consistent with the project's policies. Serving as project manager, E1 would hire the needed secretary and research associates, provide them with necessary orientation and training, coordinate the work of the involved personnel, provide the central project team with staff support, keep the project on schedule and within budget, and ensure that reports were prepared and delivered in a timely manner. The secretary would be in charge of report production and final technical editing of reports, and would have control of project information. E1's assistant would conduct the literature review and be in charge of drafting the report on planning and implementing the new PRS.

The context evaluation task group members were E2 as chair, E3, and two research associates. This group was slated to analyze the USMC's existing PRS against the adopted personnel evaluation standards and the requirements of the task order. It would report its findings to the central project team and assist the team in finalizing its report on the evaluation of the current PRS.

The input evaluation/PRS alternatives task group members were E4 as chair, E5, and a research associate. This group was assigned the task of searching out, describing, and evaluating alternative personnel evaluation systems against the selected standards and the task order requirements. It would report its findings on the state of the art reflected in alternative personnel evaluation systems to the central project team.

This proposal acknowledged that the USMC's sponsoring committee and study advisory committee and their designated representatives would provide ongoing oversight of the project team's work. Also, the proposal projected that the evaluator would deliver reports and provide in-person briefings to these committees approximately as follows:

During week 6: Final project plan

During week 10: Context evaluation of the current PRS

During week 18: Input evaluation of alternative personnel evaluation systems and literature review

During week 26: Conclusions and recommendations

During week 34: Plan for developing, implementing, and evaluating through process and product evaluations each version of the new PRS

Project Performance Plan

Building on the general plan, the proposal provided a schedule of work. The project's tasks and subtasks are listed here, followed by the scheduled period for the work. Project personnel slated to carry out each subtask are noted after each subtask.

Task 1: Organization and Background Analysis (Weeks 1 Through 6)

1. Prepare the project plan, obtain USMC approval, and choose project personnel—weeks 1 and 2: E, four days; E1, five days; secretary, two days.
2. Hire project staff—weeks 1 through 3: E1, ten days.
3. Plan for, conduct, and follow up on a meeting with the USMC's Manpower Analysis, Evaluation, and Coordination Branch to establish protocols, clarify roles and responsibilities, and present initial documentation and data requirements—during week 2: E, two days; E1, three days; secretary, one day.
4. Obtain and analyze pertinent documentation of the present PRS—weeks 2 and 3: E, two days; E1, four days; research associate, ten days; secretary, three days.
5. Prepare for, conduct, and follow up on a three-day organizational team meeting; review the project plan and PRS materials; make assignments; update the project schedule; and agree on a set of evaluation standards to recommend—weeks 3 and 4: E, five days; secretary, nine days; E1, six days; research associate, five days; E2, three days; E3, three days; E4, three days; E5, three days.
6. Prepare the first report to include an updated project plan, a recommended set of standards for judging personnel evaluation systems, and procedures and instrumentation for applying the standards—weeks 5 and 6: E, two days; E1, three days; secretary, 3 days.
7. Deliver the first report to the USMC's Evaluation Office (EO) in Washington, DC—during week 6: E, one day; E1, one day.

Task 2: Context Evaluation of the Existing System (Weeks 3 Through 10)

1. Reach agreement with the commandant by conference telephone call on the professional standards to be applied to the personnel evaluation system—about week 7: E, one day; E1, one day; E2, one day; E4, one day; E1's assistant, one day; EO representatives, one day.
2. Follow up on the conference telephone call in task 2.1 by compiling the agreed-on standards and distributing them to all participants in the project—during week 8: E, one day; E1, two days; E1's assistant, one day; secretary, one day.

3. Develop a descriptive report on how the current PRS and fitness report system are intended to operate and how they actually operate—during weeks 3 and 4: E1, three days; research associate, eight days; secretary, three days.
4. Develop a report reviewing and analyzing completed studies of the PRS and the relevant literature—during weeks 3, 4, and 5: E1, three days; research associate, four days; secretary, three days.
5. Develop a report proposing a preliminary list of performance qualities that should be measured by the USMC's unique evaluation system—during weeks 3 and 4: E1, two days; research associate, four days; secretary, two days.
6. Plan, conduct, and follow up on a two-day meeting to provide members of the context evaluation task group with an orientation and launch their evaluation of the current PRS—during week 7: E, four days; E1, four days; E2, three days; E3, four days; research associate, three days; secretary, five days.
7. Augment the literature review, prepare an updated report focused on the strengths and weaknesses of alternative personnel evaluation systems, and distribute the report—during weeks 7 and 8: E, one day; E1, two days; research associate, four days; secretary, three days.
8. Plan, conduct, and follow up on a two-day meeting to evaluate the current PRS by applying the adapted and adopted personnel evaluation standards and the USMC's requirements for the new system—during week 8: E, three days; E1, four days; E2, three days; E3, six days; research associate, three days.
9. Prepare a context evaluation report on the evaluation of the current PRS, proposing requirements to be met by the new system, and submit the report to E1—during week 10: E2, one day; E3, four days; research associate, two days.
10. Deliver a finalized context evaluation report to USMC—during week 11: E, one day; E1, one day.

Task 3: Input Evaluation to Identify and Evaluate Alternative Personnel Evaluation Systems (Weeks 9 Through 20)

1. Plan, conduct, and follow up on a two-day meeting to provide the input evaluation/PRS alternatives task group with an orientation, familiarizing them with the task 2 report—during week 10 or 11: E, three days; E1, four days; E4, three days; E5, two days; research associate, three days.
2. Prepare a report identifying, reviewing, and analyzing alternative performance evaluation systems used by other U.S. armed services, federal agencies, or appropriate civilian organizations, and submit the report to E1; gauge the appropriateness and validity of the identified systems as tools for retaining, promoting, and assigning the career force by assessing these systems against the adapted and adopted personnel evaluation standards and task order requirements for the new PRS—weeks 7 through 16: E1, one day; E4, four days; E5, three days; research associate, six days; secretary, four days.

3. Plan, conduct, and follow up on a two-day meeting of the central project team to review and reach agreements for finalizing the input evaluation report identifying and assessing alternative personnel evaluation systems and to update the plan for the remainder of the project—approximately week 12: E, three days; E1, four days; E2, two days; E3, two days; E4, two days; E5, two days; research associate, four days.
4. Finalize the input evaluation report—during weeks 17 and 18: E, one day; secretary, one day.
5. Deliver the input evaluation report to the USMC—during week 19: E, one day; E1, one day.

A unique aspect of the USMC's task order for this project was its requirement that the contractor evaluate the PRS to identify its flaws and also propose solutions. Many evaluators would resist taking responsibility for recommending solutions to identified problems. They might argue correctly that an evaluation of an evaluand can identify its strengths and weaknesses, but that such findings do not point to the best corrective actions. In this case, the evaluation team successfully addressed the issue of providing recommendations by conducting a context evaluation to diagnose problems in the USMC's personnel evaluation system and subsequently conducting an input evaluation to identify and assess alternative personnel evaluation systems that might replace the USMC's PRS. Conducting distinct but related context and input evaluations is an apt and defensible way for evaluators to address a client's request for an evaluation that both identifies problems and recommends solutions.

Task 4: Conclusions and Recommendations (Weeks 19 Through 27)

1. Plan, conduct, and follow up on a three-day meeting of the central project team to draft recommendations, including a preferred evaluation system, a field test and validation plan (including process and product evaluation designs), and a timetable for installation—during weeks 19 through 21: E, five days; E1, seven days; E2, three days; E3, three days; E4, three days; E5, three days; research associate, five days.
2. Finalize the conclusions and recommendations report, including conclusions about the reasons for the failure of the existing PRS, the merits of alternative personnel evaluation systems, and a design for a new PRS that builds on the context evaluation and input evaluation reports and meets the requirements of the task order and the adapted and adopted personnel evaluation standards—during weeks 22 through 26: E, two days; E1, two days; research associate, three days; secretary, two days.
3. Deliver the conclusions and recommendations report to the USMC—during week 27: E, one day; E1, two days; E2, one day; E4, one day.

Task 5: Planning for Development, Evaluation, and Implementation of the Proposed New PRS (Weeks 27 Through 34)

1. Meet in Washington, DC, with USMC representatives to reach agreement on steps to follow up on the conclusions and recommendations report—approximately week 27: E, one day; E1, two days.

2. Draft a plan for development, evaluation, and implementation of the new PRS—during weeks 18 through 30: E, two days; E1, seven days; research associate, ten days; secretary, three days.
3. Plan, conduct, and follow up on a two-day meeting to critique and improve the implementation plan—weeks 31 and 32: E, three days; E1, four days; E2, two days; E3, two days; E4, two days; E5, two days; secretary, four days.
4. Finalize the implementation plan for submission to the USMC—weeks 29 through 33: E, two days; E1, four days; research associate, eight days; secretary, six days.
5. Deliver the plan for development, evaluation, and implementation of the new PRS to the USMC—week 34: E, two days; E1, two days.

The initial design for this evaluation project was general. It did not specify the data collection, analysis, and reporting procedures. Instead, the evaluator's proposal stated that, once funded, the project's first task would be to produce the needed specific procedures. The project staff would develop such items as interview protocols, specifications for sampling interviewees, a plan for sampling and analyzing fitness reports, and scales and procedures for rating alternative personnel evaluation systems against the personnel evaluation standards. Given the short timeline to design and carry out the project and the evaluation team members' need to become acquainted with the USMC, it was both realistic and prudent to delay specific design decisions until the project was under way. Even in other situations in which the timeline is not short, evaluators can benefit by conducting a small planning project before committing to specific evaluation procedures. An initial "get acquainted" planning project can help the evaluators develop rapport with program stakeholders, acquire insights of use in planning the evaluation, and agree on criteria for judging the evaluand. Once the PRS evaluation project got started, the evaluator and USMC leaders agreed that the entire improvement project should be grounded in an officially adopted set of professional standards for sound personnel evaluations. This plan was based on the assumption that the project team would have approximately eight months to complete the work; thus, the thirty-four-week schedule of work. The USMC took two months to process the contract, however, and the thirty-four-week plan had to be compressed into about twenty-five weeks. Later, when the evaluator and the USMC agreed that some additional work should be done, the USMC issued a supplementary contract for about seven additional weeks of work. These developments illustrate that evaluation design often needs to be an ongoing process.

Principal Features of the Case

The evaluation design explored here was general in nature and therefore incomplete. However, it was sufficiently specific to win a $440,000 contract. The CIPP model formed the structure for this problem-solving project. The context, input, process, and product components, respectively, would investigate the following major questions: What deficiencies in the existing PRS need to be corrected? What alternative approach would meet the need for improvement best? Is the chosen new approach being carried out as intended? Is it promoting the most

deserving marines? The sponsor had not sought competitive bids, instead having chosen the evaluator for this assignment. No doubt the sponsor had found the chosen evaluator's track record and reputation to be relevant and strong and had judged that he and his team would conduct the project competently and on time and would design the details for the project. We acknowledge these idiosyncratic characteristics of the evaluation because evaluators often have to provide much more detail in their evaluation proposals than set out in the example. Clearly, each evaluation opportunity has its unique characteristics, and the evaluator should consider these when deciding how much specificity to include in the initial evaluation design. At a minimum, the prospective contractor must put forth a general methodological approach, such as this case's employment of the CIPP model, and show its relevance to the particular evaluation. In typical evaluation assignments, however, the evaluator must flesh out the evaluation design as the study unfolds.

Generic Checklist for Designing Evaluations

We now offer a generic checklist for designing evaluations (Stufflebeam, 2004a) that evaluators and their clients can use to plan the full range of relevant evaluation operations at the needed level of detail. The checklist, which appears in Exhibit 19.1, is intended as both an advance organizer and a reminder of key matters to be considered before and during an evaluation. We will illustrate the latter application of the checklist by noting how the evaluation team addressed some of its key checkpoints in the evaluation of the USMC's PRS.

Exhibit 19.1 EVALUATION DESIGN CHECKLIST

A. Focusing the Evaluation and Situational Analysis

_____ 1. Determine and clarify the evaluand and client.

_____ 2. Identify the audience and examine it vertically and horizontally. Identify the major levels and components of the evaluation audience, such as program leaders, staff, and recipients.

_____ 3. Identify audience questions, information needs, and concerns about the evaluation.

_____ 4. Identify parties who might be harmed by the evaluation, and obtain their input.

_____ 5. Examine the background of the request for the evaluation and its social and political context.

_____ 6. Identify and address the evaluation's potential barriers and other possible complicating factors—for example, the need to gather sensitive information, possible restrictions against accessing all the relevant information, human subjects institutional review requirements that may take much time to meet, requirements for confidentiality or anonymity that may be difficult to guarantee, opponents of the evaluation, prospects for misuse of

findings, prospects for nonuse of findings, conflicts of interest, issues of race and language, an indirect cost rate that may exceed what the sponsor is willing to pay, and possibly a lack of the full amount of funds needed to conduct the evaluation.

_____ 7. Identify and review relevant information, such as previous evaluations of the evaluand, evaluations of similar evaluands, pertinent literature, and relevant needs assessments.

_____ 8. Agree with the client on standards for guiding and assessing the evaluation.

_____ 9. Agree with the client on the evaluation model or approach to be applied.

_____ 10. Agree with the client on the time frame, the persons who will conduct the evaluation, key evaluation questions, required reports, client and stakeholder responsibilities, and the allowable cost for the evaluation.

_____ 11. Advise the client to fund an independent metaevaluation.

_____ 12. Decide whether to proceed with the assignment.

B. Collecting Information

_____ 1. Consider collecting a wide range of information about the evaluand: context, history, beneficiaries, benefactors, goals and structure, comparisons to similar evaluands, the schedule, resources, costs, staff, implementation, main effects, side effects, reputation, judgments by stakeholders and experts, sustainability, and transportability, for example.

_____ 2. Choose the main method for collecting information: a case study, sample survey, field experiment, or multimethod study, for example.

_____ 3. Determine the information sources: documents, files, databases, financial records, beneficiaries, staff, the funder, experts, government officials, or community interest groups, for example.

_____ 4. Determine the information collection instruments and procedures, such as interviews, participant observers, independent observers, focus groups, town hall meetings, literature review, a search of archives, the Delphi technique, surveys, rating scales, knowledge tests, debates, site visits, photography, video records, log diaries, goal-free evaluation, or case studies.

_____ 5. Specify the sampling procedures—purposive, probability, or convenience sampling—for each source.

_____ 6. Seek to address each main question with multiple methods and data points.

_____ 7. Schedule information collection, denoting times when each information source and each method will be engaged.

_____ 8. Assign responsibilities for information collection.

_____ 9. Give the client and other interested parties a rationale for the information collection plan.

_____ 10. Review the information collection plan's feasibility with the client, and consider making prudent reductions or adjustments.

C. Organizing Information

_____ 1. Develop plans and assignments for coding, verifying, filing, controlling, and retrieving information.

_____ 2. Design a database for the obtained information, including appropriate software.

_____ 3. Specify the equipment, facilities, materials, and personnel required to process and control the evaluation's information.

D. Analyzing Information

_____ 1. Identify bases for interpreting findings, such as beneficiaries' needs, objectives, standards, norms, the evaluand's previous costs and performance, costs and performance of similar evaluands, and judgments by experts and program stakeholders.

_____ 2. Specify qualitative analysis procedures—for example, thematic analysis, content analysis, summaries, scenarios, or comparisons of photographs.

_____ 3. Specify quantitative analysis procedures, such as descriptive statistics; trend analysis; cost analysis; significance tests for main effects, interactions, and simple effects; effect parameter analysis; meta-analysis; item analysis; factor analysis; regression analysis; regression discontinuity analysis; and charts, tables, and graphs.

_____ 4. Select appropriate computer programs to facilitate quantitative and qualitative analyses.

_____ 5. Plan to search for trends, patterns, and themes in the qualitative information.

_____ 6. Plan to contrast different subsets of qualitative and quantitative information to identify both corroborative and contradictory findings.

_____ 7. Plan to address each evaluative question by referencing and citing the relevant qualitative and quantitative information plus relevant alternative analyses.

_____ 8. Plan to use qualitative information to elaborate and explain quantitative findings.

_____ 9. Plan to state caveats as appropriate in consideration of any inconclusive or contradictory findings.

_____ 10. Plan to synthesize quantitative and qualitative information, for example by embedding quantitative information within a qualitative narrative or by embedding interview responses and other qualitative findings in the discussion of quantitative findings.

_____ 11. Anticipate that the client or other stakeholders may require recommendations to correct problems identified in the findings, and be prepared to explain that the same

data that uncovered the problems are unlikely to provide valid direction for solving the problems.

_____ 12. Consider planning a follow-up project to generate and validly assess alternative courses of action for solving identified problems; such a project might include an input evaluation of available alternative solution strategies, creation and evaluation of new solution strategies, engagement of relevant experts, review of relevant literature, or a working conference to chart and assess possible courses of action.

E. Reporting Information

_____ 1. Clarify the overall audience and which segments of the audience will receive which reports. For example, the program's client, staff, policy board, and beneficiaries might all receive an overall executive report, whereas particular groups might receive special reports targeted to their specific roles and interests.

_____ 2. Identify the reports needed by different audiences, such as interim, final, or component-specific reports; context, input, process, and product evaluation reports; a technical report containing specific information about the evaluation's data and procedures; an executive summary; and an internal metaevaluation report.

_____ 3. For each report, determine the appropriate format, such as printed, oral, electronic, multimedia, storytelling, pictorial, or sociodrama.

_____ 4. Outline the contents of at least the main report, showing how findings from different sources and methods will be synthesized to answer the main evaluation questions.

_____ 5. Consider dividing the final report into three subreports: program antecedents (for those who need background information), program implementation (for those who would replicate the program), and program results (for the entire audience).

_____ 6. In a technical appendix or a separate technical report, plan to include résumés of evaluation staff and consultants, information collection instruments and protocols, reports of findings from particular data collection procedures, data tables, a log of data collection activities, a list of interim reports, the evaluation contract, a summary of evaluation costs, and an internal account of how well the evaluation met the standards of the evaluation profession.

_____ 7. Develop a plan and schedule for delivering reports to the right-to-know audiences.

_____ 8. As appropriate, obtain prerelease reviews of draft reports from the client and other stakeholders.

_____ 9. Use feedback on draft reports to ensure that final versions are correct and clear.

_____ 10. Conduct feedback sessions to assist the client group in reviewing and discussing draft reports.

F. Administering the Evaluation

_____ 1. Delineate the evaluation schedule.

_____ 2. Define and plan to meet staff and resource requirements.

_____ 3. Ensure that the evaluation plan is sufficient to meet pertinent standards of the evaluation field.

_____ 4. Provide for at least internal formative and summative metaevaluations.

_____ 5. Strongly advise the client to obtain an independent metaevaluation and agree to cooperate with and supply needed information to the external metaevaluator.

_____ 6. Delineate a budget for the evaluation.

_____ 7. Negotiate an evaluation contract, specifying audiences, evaluator responsibilities and protocols, editorial and dissemination responsibility and authority, the evaluation budget, and a schedule for payments.

_____ 8. Provide for reviewing and updating the evaluation plan, budget, and contract as needed.

_____ 9. Plan for developing a stakeholder review panel and engaging this panel throughout the evaluation to review draft evaluation plans, tools, and reports and to facilitate data collection.

Source: Adapted from Stufflebeam, D. L. (2004). *Evaluation Design Checklist*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/evaldesign.pdf

As is evident in Exhibit 19.1, the logical structure of evaluation design includes elements that commonly apply to a wide range of evaluation assignments and alternative evaluation approaches. The checklist is intended as a generic guide to making decisions that typically need to be considered when planning and conducting an evaluation. The checkpoints are especially relevant when responding to a potential client's request for a demanding, complex evaluation. However, the checklist is intended for use across a broad range of evaluation assignments, both small and large, and for use with a number of approaches to evaluation. It may be used alone or in combination with other checklists. For example, it could be used with the checklists we present in subsequent chapters concerned with budgeting and contracting for evaluations and reporting evaluation findings, and the ones we reference in Chapter 25 for conducting metaevaluations. When the contemplated evaluation is small and will have only a modest budget, the evaluator and the client can find it useful to consider the full range of evaluation design issues before setting aside those that are not feasible, not particularly relevant to the situation, or not especially important. Because this checklist is intended for evaluators who work under very different circumstances and constraints, the user will need to exercise wise judgment and discretion in determining and applying its most salient parts pursuant to the needs of particular evaluations.

Although the checklist is an ordered list of elements commonly included in evaluation designs, these elements should not necessarily be addressed in a strict linear sequence. Often an evaluator cycles through the elements repeatedly while planning for and negotiating the terms of an evaluation and also during the course of the evaluation. In each such cycle, some elements are addressed, and others typically are set aside for later attention or abandoned because they do not apply to the particular situation. As noted in the example in this chapter, evaluation design is as much process as product. In using this checklist, the objective should be to develop an evaluation plan for a sound, responsive, and effective evaluation over time. We will look briefly at each section of the checklist, paying particular attention to how it applies to the evaluation of the USMC personnel evaluation system.

Focusing the Evaluation and Situational Analysis

When first considering an opportunity to conduct an evaluation, an evaluator should carefully focus the projected work to lay a sound foundation for the contemplated study. A careful preliminary investigation is also important for ensuring that it would be wise to proceed. Sometimes an evaluator will learn through early investigation and deliberations with interested parties that it is not in the cards to conduct a professionally responsible evaluation. For example, the client may want to use the evaluation to kill a program, whatever the evaluation findings. Or the client may insist on editing the final report. The evaluator should smoke out any illicit reasons for an evaluation and either obtain the needed remedies or reject the assignment. More positively, the twelve checkpoints in this part of the Evaluation Design Checklist provide a valuable guide for putting a defensible evaluation assignment on solid ground.

Essentially, the task order and evaluation design presented in the first part of this chapter satisfactorily addressed the focusing checkpoints and incorporated the up-front agreements in the evaluation plan that the USMC approved. Two exceptions can be mentioned. First, in regard to checkpoint A.6, it was not clear at the evaluation's outset that the USMC would restrict the evaluation team to interviewing only marines who were present at the Quantico headquarters. During the evaluation, it became evident that the evaluation team should interview marines on other bases, who would be more likely to give candid assessments of the PRS. At first, USMC leadership denied access to marines on other bases. Ultimately, however, this decision was reversed; the USMC issued a supplementary contract, and the evaluators obtained the needed interviews. This example illustrates the importance of negotiating matters that are vital to the evaluation's success before signing a contract. It also illustrates that design issues sometimes need to be renegotiated during a study.

A second exception to the evaluation example's having met all focusing checkpoints concerns checkpoint A.11. For whatever reason, the evaluators did not advise USMC leaders to fund an independent metaevaluation. Essentially, the two USMC panels provided this function by reviewing all of the evaluation reports. Ultimately, the USMC commandant provided the evaluators with a unit citation for outstanding service to the USMC. The evaluators were proud to accept this judgment as one kind of important independent metaevaluation. Nevertheless, one could fault the evaluators for not specifically recommending that the USMC fund an independent metaevaluation of the evaluation team's work and reports.

Collecting Information

The second category of checkpoints in Exhibit 19.1 deals with the core issue of collecting information from which to judge an evaluand. Checkpoint B.1 advises the evaluator to consider collecting a wide range of information, such as the evaluand's background, structure, activities, costs, other resources, and outcomes. To obtain the selected information, checkpoint B.2 calls for choice of an appropriate information collection framework, which often will entail a combination of methods. Checkpoints B.3 through B.6 ask for details concerning information sources, instruments and methods, and sampling procedures and how information will be combined to answer each evaluative question. Checkpoints B.7 and B.8, respectively, require developing an information collection schedule and assigning responsibilities for collecting the information. The final two checkpoints in this category involve justifying the information collection plan and considering whether it should be reduced or adjusted. The information collection checkpoints can all be considered when first planning an evaluation. Realistically, however, decisions on many of these matters often are made as an evaluation proceeds, and even then may have to be revised later.

The evaluation team in the military evaluation example initially addressed the information collection checkpoints by employing the CIPP model, which calls for collecting a wide range of information through context, input, process, and product evaluations. Once the evaluation was funded, the context evaluation and input/PRS alternatives evaluation task groups developed and implemented specific data collection plans.

The context evaluation task group pursued a multimethod approach. Among the methods employed were the following:

- Content analysis of past evaluations of the PRS
- Examination of recorded problems, recommended improvements, and subsequent actions drawn from previous evaluations of the PRS and from a succession of action reports
- Content analysis of PRS regulations, procedures, and forms
- Content analysis of the code phrases in an unofficial guide, commonly known to and used by USMC supervisors, for writing fitness reports
- Preparation and use of interview guides to obtain information from a cross-section of marines and members of promotion boards
- Mailing a survey to a representative sample of sergeants being reviewed for promotion
- Obtaining promotion records and analyzing them in terms of ranks, military specialties, gender, race, and venues of service
- Computer-based content analysis of a sample of seventy-five thousand fitness reports and results of promotion panel assessments
- Observation of sessions to train supervisors in evaluating subordinates
- Construction and application of a scale for rating the PRS against the USMC's adapted and adopted personnel evaluation standards and for listing strengths and weaknesses for each standard

The input evaluation/PRS alternatives task group implemented a two-stage study. First, it conducted case studies of the personnel evaluation systems of four U.S. military services, two foreign military services, and two U.S. private corporations. It subsequently rated and listed strengths and weaknesses of each system against the USMC's adopted personnel evaluation standards. Because none of the reviewed systems satisfactorily met the twenty-one applicable standards, the input evaluation team next conducted an advocate teams study. This first entailed engaging three teams to study the information so far amassed and to use it to create three competitive proposals for a new PRS. Once these proposals were generated, the evaluation's central project team rated each one, listed its strengths and weaknesses against each of the twenty-one applicable standards, and reported the findings to USMC's sponsoring committee.

Organizing Information

Evaluations require an effective approach to information management. For each set of information, an evaluator needs to follow systematic steps to ensure its accuracy and security. Some data must be coded for later summary and analysis, and some will need to be keyed into a computer. In all such operations, the evaluator should train those who will carry out the work, supervise them, regularly check their work for accuracy, and ensure that only authorized personnel access and use the stored information. As seen in checkpoint C.3, after developing an information management plan, the evaluator should arrange for the equipment, facilities, materials, and personnel needed to process and control the evaluation's information.

Early in an evaluation, the evaluator needs to establish a functional system to file, control, and retrieve information that directly reflects the evaluation's structure. Although no one system of categories of collected information and other evaluation materials would apply to all evaluations, general examples can be offered. Materials involved in focusing the evaluation could include the task order or RFP; the proposal; the contract; the budget; human subjects institutional review records; staff; consultants; evaluative standards and criteria; correspondence folders for key participants; pertinent background reports and literature; the evaluation schedule; and rules for accessing, using, and returning filed information. Information collection files could be divided into such categories as methods and data collection instruments, sampling plans, information sources and their protocols, information collection assignments, plans and materials for training information collection personnel, news clippings, and specific data collection schedules. Information analysis materials could include analysis plans for all sets of information, plans for synthesizing findings from different sets of information, and information analysis assignments. The category of reports and reporting actions could include draft and final versions of all reports, records of stakeholder critiques of draft reports, technical appendices, multimedia materials to support presentations of findings, and plans for presenting findings.

A functional evaluation project filing system should have clear rules and arrangements for keeping the information secure, while giving evaluation team members ready access to pertinent information. The lead evaluator should establish a list of personnel who can access

and use the information. In many evaluations, the evaluator should remove or have removed the identities of individuals associated with given evaluation records before making them available to evaluation team members for review and analysis. Evaluation project files should be maintained in locked filing cabinets in lockable offices and be controlled by an evaluation team member, such as the project secretary. It is also a good idea to have a practice in place of signing a log sheet before checking out a piece of information. In general, the rules and procedures of any good library apply to the control and use of evaluation project records.

In some evaluations it is appropriate to establish and maintain a database, especially if one is to track and record the performance of an evaluand over time. The establishment and use of such a database require selection of appropriate computer software and a process and assignments for checking, coding, verifying, and recording data. The evaluator needs to ensure that those who are tasked with carrying out the database functions are appropriately trained and supervised. Again, provisions should be made to keep such information secure and accessible only to authorized persons.

Analyzing Information

This section of the checklist provides detailed suggestions for analyzing evaluative information. These suggestions basically are self-explanatory, and we will make only some supporting general observations about analysis issues. Further discussion and specific examples are provided in Chapter 23.

In general, analyses in an evaluation should be keyed to answering the basic evaluation questions and judging the evaluand. As seen in checkpoint D.1, optional bases for judging a program's merit and worth include assessed needs of beneficiaries, program objectives, professional standards (as were employed in this chapter's military evaluation case), national or state norms, a previous level of performance, performance by similar evaluands, and judgments of the evaluand rendered by pertinent experts or beneficiaries. Together the evaluator and client should determine the bases that will be most appropriate in the particular study.

To develop and support judgments of an evaluand in relation to the selected bases, the evaluator needs to employ systematic analyses of both qualitative and quantitative information. The technical literature of research and evaluation contains a rich cornucopia of pertinent methods, as is evident in checkpoints D.2 and D.3. There also exists a wide range of relevant computer programs that are available for purchase or even for free. The evaluator needs to choose methods and software that are appropriate for conducting the data and information analyses needed to answer the key evaluation questions. In choosing analysis methods, the evaluator should make sure that the assumptions required to apply these methods can be substantially met by the available information. Often it will not be possible to meet a procedure's required assumptions perfectly concerning such matters as the nature of the measurement scale, randomization of subjects, and independence of observations. In such cases, it can be especially useful to apply multiple analysis procedures to a given data set to provide checks and balances on findings from the different analysis techniques. In general, it is good analysis practice to contrast different subsets of qualitative and quantitative information to identify both corroborative and contradictory findings.

Ultimately the evaluator needs to synthesize results from analyses of the different sets of information. The objectives of the synthesis are to combine findings to answer each evaluation question and to reach bottom-line judgments of the evaluand. In presenting these judgments, we think it is useful to organize them into conclusions about the evaluand's merit or quality, its worth in addressing beneficiaries' needs, its superiority to other objects based on cost-effectiveness, its significance for use in other settings, and its probity as an ethical response to the beneficiaries' needs.

Some final words are in order in regard to recommendations. In typical evaluations, the resultant information and analyses do not provide justifiable bases for making recommendations on how to improve or, especially, replace the evaluand. The reason is that the evaluator has not gathered and analyzed information about such matters as the merit, worth, and cost-effectiveness of the contemplated recommendations. Typically the recommendations are conceptualized or identified through group deliberations, but are not grounded in empirical study. It can be misleading and professionally irresponsible to advance recommendations that rest on shaky ground. An exception is seen, however, in the input evaluation component of the CIPP model, which can be employed in helping design a program or other intervention (as in the USMC case) or as a follow-up step after a program has been fully evaluated. In a follow-up input evaluation, an evaluator would systematically generate and validly assess alternative courses of action either for solving problems that were identified by the original evaluation or for replacing the evaluand. Under such circumstances, the evaluator stands on solid ground when presenting recommendations, because they are based on systematic, comparative, empirical inquiry. Of course, to generate an empirically grounded recommendation, the evaluator needs to come to an agreement with the client that such a follow-up investigation of optional courses of action is needed and will be appropriately scheduled and funded.

Reporting Information

A fundamental goal of any evaluation is to communicate findings effectively to members of the audience and secure their appropriate use of the reported information. Steps to promote and secure use of findings are critical parts of the evaluation design and need to be addressed at the outset of planning the evaluation and throughout the study. The identity of the main client will be clear at the outset. For example, in the USMC personnel evaluation case, the key client was the USMC's commandant. But an influential evaluation should address the questions and information needs of a much broader group than just the client. It must reach those who will make decisions based on the findings, those who will have operational responsibility for applying the findings to improve the evaluand, those who are paying for or using the assessed program, and those who are the subjects or recipients of the program.

It is vital to study all segments of the intended audience to identify their different information needs and prepare and schedule delivery of appropriate reports. In general, the different reports might include component-specific reports, such as context, input, process, and product evaluation reports; reports keyed to particular methods, such as surveys, ratings,

case studies, or content analysis; interim progress reports; the final report; and a technical appendix or a separate technical report. Depending on the needs of the audience for each report, the evaluator might plan to employ a variety of formats beyond the printed report. Presentation modes could be oral, electronic, multimedia, focus group, town hall meeting, sociodrama, or webinar approaches.

The evaluator usually should arrange to have appropriate interactions with the client and others throughout the evaluation. This is important to discern their most important information needs, motivate them to receive and use the findings, deliver evaluation findings when they can best be used, obtain their assistance in gathering data, and receive stakeholders' critiques of previous reports. To promote use of findings, the evaluator should seize appropriate opportunities to engage members of the audience in exchanges about evaluation plans and findings. In advance of data collection, it can be useful to outline the contents of a projected report, complete with "dummy tables," and to go over this with the client or other audience members. In later exchanges, the evaluator should seek explicit feedback from the audience about the strengths and weaknesses of previous reports and about what information would be most useful in future reports.

Some cautionary notes are in order. One is that the evaluator should take care to maintain the evaluation's independence while obtaining feedback from the full range of stakeholders. The evaluator should be open to receiving feedback from all interested parties, carefully assess the relevance and dependability of such feedback, and use it as appropriate in generating evaluation findings. Also, the evaluator should not empower stakeholders to make or strongly influence evaluation design decisions and must not pander to any illegitimate stakeholder desires and interests.

Often interim reports are as influential as, or even more influential than, the final report. This is especially so in formative, decision-oriented, and responsive evaluations. In such studies, the evaluator should plan carefully and carry out the interim evaluation reporting effectively. As mentioned in checkpoint F.9, it can be very useful to establish a review panel representing the different levels and segments of the audience, periodically deliver findings to this group, interact with the group about the relevance of the reported information, obtain the group's critical reactions to each report, and obtain their views about what information would be most useful in future reports. Such a panel can also be a useful source of assistance in realistically scheduling and facilitating the data collection process.

Clients typically require, beyond the interim reports, a comprehensive final report. We have found it useful to divide such a report into three main parts. A program antecedents part can be useful to persons who need background information on the program, including when, why, and how it was started and by whom; its location and environment; and its institutional home. The program implementation part can be of special interest to groups that might want to replicate the subject program. This part should be highly descriptive rather than evaluative. It should identify the program's objectives, beneficiaries, governance, staff, organization, operations, and funding. The program results part should be addressed to the entire need-to-know audience. It should summarize the evaluation design and process; present the findings for each main question; and synthesize the findings to present conclusions in regard to the program's merit,

worth, significance, cost-effectiveness, sustainability, transportability (as appropriate), and probity. Further, a final evaluation report should include, in addition to these three main parts, a technical appendix or a separate technical report. As seen in checkpoint E.6, this technical portion could include résumés for all members of the evaluation team, information collection instruments and protocols, reports of findings from particular evaluation procedures, data tables, a log of evaluation activities, a list of evaluation reports, a summary of evaluation costs, a copy of the evaluation contract, and an internal account of how well the evaluation met the standards of the evaluation profession.

Lessons learned from the USMC personnel evaluation suggest that the utility of interim reports is enhanced by conducting review sessions and carrying out steps such as the following:

1. Engage the client to appoint a review panel that is representative of program stakeholders.
2. Secure the client's agreement to chair the review panel, and make panelist responsibilities clear.
3. Schedule each feedback session with the panel well in advance.
4. Distribute the most recent draft report along with an agenda to the review panel members about ten days prior to a given feedback session.
5. Determine with the client the agenda, location, and time frame for the session.
6. Have the client start the session by going over the session's objectives and agenda.
7. Use a multimedia approach to brief the review panel on key aspects of the report, including questions, methodology, obtained findings, and key issues for discussion.
8. Engage the review panel's chair to lead a discussion of findings and their implications for action.
9. Assist in the panel's discussion of findings as appropriate, but do not dominate or become defensive.
10. Following the chair's discussion of findings with the panel, ask panelists to voice their reactions to the report; identify their most important information needs for future reports; and, as appropriate, assist future data collection efforts.
11. Have the chair ask each panel member to cite the meeting's most important outcomes and then summarize the meeting from his or her perspective.
12. Schedule the next review session, summarize pertinent next steps, and thank all present for their participation.
13. Following the meeting, prepare the minutes and distribute them to all meeting participants.

Administering the Evaluation

All evaluations require effective administration, a key responsibility of the lead evaluator. The initial evaluation plan should include a schedule of evaluation activities and staff assignments, which should be updated as appropriate during the evaluation. The schedule should be worked out with the client to ensure stakeholder availability in data collection as well as

reporting. In consideration of the scheduled evaluation tasks, the plan should include an appropriate budget. Among the key cost items are evaluation staff, consultants, materials and equipment, facilities, communication and other services, travel, and indirect costs. Building on the evaluation schedule and budget, the evaluator also should negotiate a contract that guarantees the evaluation's viability and integrity. The evaluation plan should provide for reviewing and updating the evaluation design and contract as needed. Later chapters address the budgeting and contracting tasks in considerable detail.

A very important administrative task is staffing the evaluation. Here the evaluator should recruit, assign, train, and coordinate staff members such that they effectively carry out all aspects of the evaluation and earn the confidence of the client and other stakeholders. Required competencies often include high-level skills in measurement, statistics, and computer technology; in-depth knowledge of qualitative methods; the ability to establish rapport and working relationships with personnel in the field; knowledge of the evaluand's content; facility with multimedia presentation methods; and excellent writing, editing, and oral communication skills. In developing an evaluation team, the lead evaluator should take into account the ethnic and other characteristics of stakeholders and ensure that the assembled evaluation team can earn the trust and confidence of all segments of the audience. It is often desirable for the evaluator and client to arrange for the involvement of a review panel whose members are representative of the program's stakeholders.

Finally, the evaluator should take steps to ensure that the evaluation will meet the standards of the evaluation field. The evaluation design should be grounded in appropriate standards, and the evaluator needs to obtain the client's endorsement of the standards. In addition, the evaluator should provide internal formative and summative metaevaluations of the evaluation work in light of the adopted standards and advise the client to contract for an independent metaevaluation of the completed study. We address this topic in depth in Chapter 25.

Summary

In this chapter we have provided a perspective on the crucial topic of evaluation design. We began by summarizing and discussing an actual design based on the CIPP model. Although this design was quite general, it was funded at a level of $440,000, and its implementation led to a highly influential and well-received evaluation. This case illustrates that evaluation design is as much process as product, as the design for evaluating the USMC personnel evaluation system had to evolve and broaden along the way.

We then presented a generic checklist for designing evaluations. It is configured to work not only with the CIPP model but also with any other defensible evaluation approach. Moreover, it is intended to be useful in generating an initial design and periodically reviewing and updating that design. It is intended not as a cookie-cutter approach to design, but as a flexible tool that affords evaluators of both large and small studies ample room for selectivity and creativity. Once an evaluator has worked out a design that is responsive to and properly reflective of a client's evaluation needs, we think he or she would find the checklist of further use for guiding

the evaluation and controlling its quality, and ultimately for assessing and reporting the extent to which all aspects of a sound and functional evaluation were addressed.

REVIEW QUESTIONS

1. From reading this chapter, how would you define evaluation design? In general, what main topics should be included in an initial evaluation design? Why must the evaluator make design decisions before the evaluation commences? What skills are required to develop sound evaluation designs?
2. "Evaluation design is both process and product." Discuss this statement in the context of designs for preordinate evaluations and responsive evaluations.
3. In general, what roles are involved in executing an evaluation design, and for what skills and in what ways should the design provide for meeting the training needs associated with each of these roles?
4. It is important to restate a potential client's criteria for evaluating the evaluand. Such a summary serves several useful purposes. What are some of these?
5. In the project discussed in this chapter, the military organization agreed to the evaluator's recommended set of standards for the new PRS but added a standard—Transition to the New PRS. Give reasons why this was an important addition.
6. When helping develop a large-scale personnel evaluation system such as the one depicted in this chapter, the evaluation team should carry out a succession of tasks for producing and validating the new system. Within the context of the CIPP model, list and provide rationales for at least five such tasks.
7. Conducting distinct but related context and input evaluations, as we have stated, is an apt way for evaluators to address a client's request for an evaluation that both identifies problems and recommends solutions. Give reasons why this approach is defensible, bearing in mind that it is usually inappropriate for an evaluator to offer the best corrective actions. Also, explain why typically it is problematic for evaluators to present recommendations.
8. The importance of negotiating features of the evaluation before signing a contract based on an agreed-on design cannot be overestimated. Why is this so?
9. If it is to be influential, why should an evaluation address the questions and information needs of a considerably broader group than the client? How does consideration of this matter have a bearing on the design of the evaluation?
10. According to this chapter, what are the roles of an evaluation review panel? How should such a panel be engaged during the course of an evaluation? What are the risks in employing the review panel, and how can the risks be avoided?

Group Exercises

Exercise 1

The board of a nationwide retailer whose sales have slumped over the past five years has decided to have the company evaluated. In brief, the RFP states that the focus will be on national and state administrations, management of individual stores, and communication among these entities. The RFP gives the green light for the evaluator to examine such peripheral issues as quality control and deployment of stock. Moreover, if other major deficiencies in the system become apparent during the evaluation, these also will become part of the study following consultation with the board chair, the managing director, and other persons on the review committee. Initially, an evaluation project performance plan is required from respondents to the RFP to indicate the practicality and quality of a schedule of work that the board could expect from each respondent. The RFP specifies a maximum funding level of $600,000 for the evaluation, a requirement for a fixed-price contract, and a deadline for delivering the final report of not more than three years after negotiation of the evaluation contract.

Imagine that your study group constitutes a sizable evaluation firm. Realizing that the requested undertaking would be a huge task, requiring detailed planning and a considerably expanded and skilled evaluation workforce, you nevertheless decide to respond to the RFP. Using as guides this chapter's Evaluation Design Checklist and the schedule of work outlined for the military evaluation example, develop a possible schedule of work (providing only general consideration at this early stage of time allotments for evaluation tasks). One member of your group could act as a recorder of points developed under each of the six categories of checkpoints in the Evaluation Design Checklist. For an example of how another evaluation team responded to a similarly large evaluation assignment, your group should find it useful to refer to the schedule of work that was outlined in this chapter for evaluating the USMC's PRS.

Exercise 2

This chapter has emphasized that planning an evaluation design is an ongoing process. Thus, your initial plan developed in exercise 1 will necessarily be general. Describe the process your group would follow to make the design more specific, concrete, and actionable.

Exercise 3

Return to your evaluation project performance (tasks) plan, and discuss whether or not your group should add language to this plan regarding the possible need for changes in evaluation procedures and funding as the evaluation unfolds. As the evaluation study develops, it is reasonable to assume that the client group might later pose new questions to the evaluation team. Discuss and decide whether your group should (1) include language in the initial evaluation plan to ensure a reasonable level of flexibility and a process for adjusting procedural plans and modifying the budget as the evaluation develops, or (2) withhold such consideration until and if it becomes clear that evaluation procedures and/or funding have to be changed.

In discussing this issue, consider whether building flexibility into your evaluation plan would jeopardize your group's chances of winning the evaluation contract or, instead, add credibility to the plan.

List the criteria that your group used to decide whether or not to add language on the need for flexibility to your evaluation plan.

If you decided as a group to build some flexibility into your evaluation plan, write a brief passage that your group might insert into the plan for that purpose. If you decided not to make any mention of the need for flexibility in your evaluation plan, write a brief justification for this decision.

Exercise 4

Obtain and read the report of a completed evaluation. Then, as a group, employ this chapter's Evaluation Design Checklist to identify the extent to which all checkpoints were effectively addressed. Discuss whether the checkpoints not addressed were not actually applicable, or whether some or all of them should have been addressed. Subsequently, discuss the adequacy of the evaluation's design, using the checklist as a reference.

Suggested Supplemental Readings

Davidson, E. J. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation*. Thousand Oaks, CA: Sage.

Joint Committee on Standards for Educational Evaluation. (1988). *The personnel evaluation standards*. Thousand Oaks, CA: Corwin Press.

Stufflebeam, D. L. (2004). *Evaluation Design Checklist*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/evaldesign.pdf

CHAPTER 20

BUDGETING EVALUATIONS

This chapter provides a rationale and ethical principles for evaluation budgeting, an illustrative evaluation budget, identification and discussion of the various dimensions of evaluation budgeting, a detailed evaluation budgeting checklist, and step-by-step guidance for building evaluation budgets.

While developing an evaluation design, the evaluator cannot escape bearing in mind costs associated with the planned study. These need not be explicated in any detail initially, but inevitably costs will become an integral part of planning an evaluation. Evaluation design and budgeting are two early basics of planning an evaluation. A budget should provide a best estimate of the funds required to successfully carry out the full range of planned evaluation tasks. The design will have indicated most if not all of the tasks to be performed; an analysis of these will give an indication of predictable costs. The structure and specificity of initial evaluation budgets will vary, however, depending on the nature of the evaluation project and the type of financial award.

In 1980 Cronbach and Associates stated that "deciding on a suitable level of expenditure is . . . one of the subtlest aspects of evaluation planning" (p. 265). This observation could suggest that a budget ideally could be developed and agreed to through consultation between evaluator and sponsor. The ideal situation, unfortunately, is not the norm. Collaborative planning may not be possible from the beginning if the sponsor sets budgetary limitations for the study. Such a situation can be frustrating for an evaluator, particularly an experienced one, because his or her ability to assess budgetary requirements is set aside. More frequently, however, the sponsor does not have a

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **The logic underlying evaluation budgeting, including the interplay between evaluation budgeting and evaluation design**
- **Ethical imperatives in budgeting evaluations, plus unethical budgeting practices to steadfastly avoid**
- **Details of an illustrative evaluation budget**
- **Budgeting under different types of evaluation agreements: grants, contracts, and cooperative agreements**
- **Different types of evaluation budgets: fixed-price, cost-reimbursable, and cost-plus budgets**
- **Different schemes for breaking out budgets: line items, tasks, time periods, and various combinations**
- **Evaluation budgeting for preordinate versus responsive evaluation approaches**
- **Main tasks in building an evaluation budget**

definite sum of money (initially at least) in mind, and, in general, the development of a budget is left in the evaluator's hands. Whether this budget is accepted by the sponsor depends strongly on how it is stated, developed, and justified. In this chapter we present ways to formulate a convincing, ethical, and defensible budget.

Although a sound and detailed evaluation design provides a solid grounding for budgetary discussions and possible decisions, a sponsor nevertheless may be quite unaware of the extent of information an evaluation may produce and also of attendant costs. If evaluator-sponsor discussions clarify such issues, including budgetary limitations, at an early stage of the evaluation, the chances of effective decisions at later stages of the study are enhanced. As we pointed out in Chapter 19, evaluation designs must remain sensibly flexible. Similarly, if at all possible, budgets must retain a degree of flexibility, particularly if new opportunities, which could be exploited, arise as a study progresses. We offer advice on this matter later in this chapter when discussing budgetary cushioning.

The budgeting process should be sensitive to whether the evaluation will be preordinate or responsive. The evaluation tasks for preordinate evaluations are delineated in advance, and an up-front budget can be specified in detail and considered relatively fixed. In contrast, budgets for responsive evaluations necessarily are general and in need of periodic updating as sponsors respond to interim reports and update their information requirements. Sponsors will differ in the amount of specificity they require, depending on the type of award: grant, contract, or cooperative agreement. Under a grant, the sponsor mainly wants to ensure that the evaluation will produce informative, high-quality outcomes and usually will not ask for extensive details on how funds will be allocated and expended to achieve the result. Under a contract or a cooperative agreement, the sponsor is more likely to require detailed breakouts of projected charges. Another consideration in the budgeting process is that the sponsor may be uncertain of which evaluation components it is able and willing to support. Accordingly, the sponsor might ask for a modular budget, allowing for funding only part of the evaluation or withholding decisions about certain components until later. An added consideration is the evaluator's institution, which will require a certain level of detail for internal accounting and auditing purposes, regardless of the sponsor's requirements for specificity.

We begin by providing an overall perspective on the ethics of building and implementing evaluation budgets. Then we report and discuss the budget for the military evaluation example presented in Chapter 19. Because that evaluation employed a fixed-price contract and was quite general, we also examine budgeting under other types of agreements. Finally, we present and explain a generic checklist for building an evaluation budget.

Ethical Imperatives in Budgeting Evaluations

Basically, an evaluation budget should enable the evaluator to implement the full range of proposed tasks at such a high level of quality and professionalism that the sponsor is convinced of the potential and defensible strength of the study. If the sponsor's funding constraints preclude successfully carrying out the requested evaluation tasks, the evaluator

should consider pursuing the following actions. The first is to justify and request the additional funds needed. If the sponsor cannot agree to increase funding, the evaluator should consider seeking a reduction in the study's scope to do well whatever is done with the available funds. Failing such a scope reduction effort, the evaluator should consider respectfully declining the assignment to avoid doing a marginal- or low-quality piece of work. If an underfunded assignment cannot be declined, the evaluator should consider stating in the final evaluation report that funding restrictions might have negatively affected the reliability and validity of reported findings.

The main point here is to emphasize that before proceeding with a planned evaluation, the evaluator should take all reasonable steps to ensure that he or she will have the resources necessary to conduct a professionally responsible study. Such assurances are obtained through up-front agreements on the evaluation tasks and funds. Alternatively, the evaluator might get the sponsor to agree in writing to clarify information needs along the way and allocate additional funds as appropriate. Successful evaluation budgeting is not simply about winning evaluation grants and contracts; it is about establishing and maintaining financial viability and fiscal integrity for a professionally defensible evaluation.

We cannot overemphasize that it is wise and professionally responsible to ground a designed evaluation in an up-front, honest, and competently prepared estimate of the needed funds. To the extent that the evaluation assignment is fully designed in advance, the evaluator should provide his or her best estimate of a full-cost budget. If the evaluation assignment is general and expected to evolve, the evaluator should offer a tentative budget aligned with what is known of the assignment and include provisions for periodically reviewing and updating the budget as the evaluation assignment develops. The general principle is that evaluators should build budgets that fairly and accurately present the level of funds needed to professionally carry out the designed evaluation activities.

An evaluator should not act dishonestly or incompetently in budgeting a designed evaluation. Unfortunately, it is not uncommon for evaluation contractors to submit highly inflated budgets or gross underestimates of the needed funds. Overestimates of needed funds occur especially in sole-source situations in which an evaluator initiated a proposal to conduct an evaluation or in which the sponsor sought services from a particular evaluator. In either case, an evaluator who is bidding on a noncompetitive award might intentionally grossly overbid a job to make a sizable profit or do so out of inexperience and incompetence. We see opportunistic, sponsor-gouging practices as unprofessional. Any possibility for inflating an evaluation budget is a prime reason why a sponsor should seek an independent assessment of a prospective evaluator's evaluation design and budget.

A fringe area of overbudgeting has caused some sponsors to be cynical about hiring university professors to conduct evaluations. We have seen all too many cases in which the evaluator surreptitiously inflated an evaluation budget to support activities with no relationship to the evaluation. The excess funds were used for such unauthorized purposes as supporting graduate students, conducting research, hiring outside speakers, paying employees who did not work on the evaluation, buying equipment or furniture not related to the evaluation, or funding trips to conventions. An evaluator can justify such expenditures of contracted

evaluation funds only if the sponsor previously agreed that the expenditures are an acceptable part of the contracted budget. Otherwise, we believe the evaluator would be misappropriating the evaluation funds and engaging in professional misconduct by using contracted evaluation funds for unauthorized uses. As part of an up-front negotiation, the sponsor might agree to a cost-plus contract to allow support of functions not related to the evaluation. The sponsor could see such overinvestment as part of the price of obtaining the evaluator's services. A cost-plus budget is appropriate as long as the evaluator clearly discloses the elements that are unrelated to the evaluation and the sponsor agrees to fund them.

In the case of underbidding a job, a naive, do-gooder, neophyte evaluator might unintentionally request far less than the needed funds and later fail in the assignment or face the embarrassment of admitting poor budgetary planning and requesting additional funds. In a competitive bidding process, an experienced but unscrupulous evaluator might intentionally submit a lowball budget to win the competition and thus essentially buy the contract. In such cases, the contractor knows beforehand that he or she will have to return to the sponsor lamenting that the original budget unfortunately was an underestimate and that additional funds are needed. In cases where the evaluator grossly underbid the evaluation job, intentionally or unintentionally, and got funded, the sponsor faces three undesirable options: give in to the evaluator's request for more money, let the evaluation proceed with insufficient funds and produce poor outcomes, or cancel the contract and incur the loss of expended funds without receiving the needed report.

We advise evaluators to act professionally and prudently in preparing their budgets. When possible, evaluators should develop their best estimate of a full-cost evaluation budget. In this way, they deal honestly with the sponsor and also ensure that the evaluation will have the necessary funds to succeed. However, there is bound to be error in all budget projections for an evaluation. Accordingly, evaluators sometimes need to err on the side of requesting slightly more funds than might be needed. Such padding of the budget is in the interest of ensuring that the evaluator can cover unexpected costs and produce a high-quality evaluation. Moreover, if an evaluation concludes with a budgetary surplus, the evaluator can and often should offer to return unused funds to the sponsor.

A full-cost budget with no inflation factor for contingencies might still prove to be larger than necessary. This can occur for a variety of reasons. For example, not all needed staff may be hired as soon as was projected, thus saving on personnel costs. Or some projected participants in evaluation meetings might not make all the needed trips, thus saving personnel and travel money. It might prove possible to institute cost-saving measures and thus expend less for the evaluation than was originally projected. For example, conference telephone calls might be substituted for originally projected face-to-face meetings. Or if the evaluator is conducting one or more other projects in the study's geographical area, travel costs might be split between the different projects, thus saving money for all of them.

Given the imperatives to conduct a sound evaluation and also to ask only for fair financial compensation, the evaluator should take appropriate steps not to overcharge the sponsor. Although an initial full-cost budget helps ensure that an evaluation can succeed, we also advise the evaluator to be frugal in the interest of making the evaluation cost effective. If it is

allowed or required by the contract, the evaluator should consider returning unused funds to the sponsor or possibly agree to use the surplus to conduct additional relevant work. In our experience, however, sponsors do not always want or have authority to accept returned funds or additional services. Nevertheless, it is ethical for the evaluator to discuss these possibilities with the sponsor.

We have sought to project a strong ethical position pertaining to evaluation budgeting. However, some caveats should be reiterated. The advice to prepare an up-front, full-cost budget is particularly appropriate to preordinate, extensively planned evaluation studies. In such cases, one knows essentially what will occur in the evaluation and can make quite reliable cost estimates. However, costing out evaluation work is less tractable in responsive evaluations in which information needs will continually unfold. In such cases, we advise the evaluator to provide best estimates of evaluation costs in the original budget, along with stipulations that the cost estimates will be revisited and updated to keep pace with evolving evaluation requirements. Alternatively, the evaluator might request a small planning grant to allow development of a sound evaluation design and budget prior to contracting for the entire evaluation. In all efforts to build an evaluation budget, the evaluator should strive to budget for what is required to produce a professionally responsible evaluation, exercise utmost fiscal integrity, and disclose any requests for funds beyond those necessary to conduct the study.

Fixed-Price Budget for Evaluating a Personnel Evaluation System

The evaluation of the military personnel evaluation system presented in Chapter 19 was grounded in a fixed-price, sole-source contract. The sponsor stipulated that the evaluator had to complete the project within eight months and could request whatever money would be necessary to complete the job well and on time. (The evaluator thus requested and received $431,763 to support the evaluation.) Moreover, the sponsor stressed that the evaluator should be sure to request all the needed funds because there would be no opportunity to renegotiate this amount.[1] An evaluation plan and budget were needed almost immediately. With all due haste, the evaluator prepared and submitted the budget that follows.

Considering the relatively large amount of requested funds, the budget in Table 20.1 includes little detail. Under a fixed-price agreement, the sponsor found the level of specificity sufficient, because the evaluation team was contractually obligated to complete the stipulated tasks to the satisfaction of the military organization by the mandated deadline and because cost was not an issue of primary importance to the sponsor.

However, the detail in this budget was not sufficient for the evaluation team's parent organization. Its administrators needed considerably more specific information to ensure that resources would be sufficient to do the job, to charge project expenses to the proper line items, to account for expenditures, and to audit the effort. The evaluation team was able to meet institutional budgeting requirements because of the way the evaluation design was constructed. Accordingly, they provided their budget office with budget notes that explained projected costs for personnel, fringe benefits, travel, consultants, supplies, services, and indirect costs.

Table 20.1 Budget for the Project to Evaluate the U.S. Marine Corps Personnel Evaluation System

Line Item	Cost	Total
A. Personnel		
1. Salaries	$98,284	
2. Fringe benefits	$38,331	
3. Total personnel		$136,615
B. Travel		$20,439
C. Consultants		
4. Honoraria	$82,000	
5. Travel plus support services for consultant team leaders	$65,526	
6. Total consultants		$147,526
D. Supplies		$1,760
E. Services (telephone, photocopying, and postage)		$27,444
F. Total direct costs		$333,784
G. Total indirect costs		$97,979
H. Total project costs		$431,763

Personnel costs were derived and delineated from the work plan in the evaluation design. That plan was explicit in noting the number of days each staff member would work on the project. A basic personnel cost was determined for each staff member by multiplying his or her daily rate by the number of days to be worked in the project. This provided the institution with the basic line-item estimated cost for each staff member who would work on the evaluation. In preparing the budget for the sponsor, the evaluation contractor had summed the cost for each staff member to obtain the salary total of $98,284. The $38,331 for fringe benefits was determined by multiplying the institution's 39 percent fringe rate by the total salary amount ($98,284). The total estimated personnel costs then came to $136,615.

The estimated travel costs for the institution's staff were also built from the evaluation design. It was estimated that costs for each trip by a staff member would average $650 for plane tickets and rental cars and $160 per day for lodging, meal, and associated expenses. Because the staff members were projected to make twenty-one trips involving thirty-seven days, the total travel cost for staff amounted to $19,570, plus a (padded) amount of $869 for incidental and unexpected meeting expenses. These amounts were within the institution's guidelines for travel expenses in expensive venues. The amounts summed to the staff travel total of $20,439.

The consultant honoraria figure was determined by summing the number of consultant days in the evaluation design document (seventy-six), adding six to allow for the hiring of additional consultants if needed, and multiplying the total of eighty-two by a daily consultant rate of $1,000. This yielded the total consultant honoraria figure of $82,000.

The consultant travel plus support services for consultant team leaders figure was determined by totaling the number of consultant trips found in the evaluation design (twenty-five); identifying, from the design, the total number of days the consultants would be engaged in

the trips (sixty-seven); multiplying twenty-five by $750; multiplying sixty-seven by $200; and summing the two products. To this result of $32,150 the evaluator added $33,376. Each context and input evaluation consultant team leader was allotted $15,000 for support services and their discretionary use. In addition, $3,376 was allocated to a contingency fund for paying the travel costs of additional consultants that probably would need to be hired. These budgetary provisions accounted for the consultant travel plus support services line item of $65,526.

It is noteworthy that the travel cost rates used for staff were lower than those for the evaluation's consultants. This was due to such circumstances as all staff members' being located at one site relatively close to the location of most of the projected off-campus work and using the same airline and car rental companies. Also, the institution could negotiate airplane and rental car costs for staff in advance. The more varied and less predictable travel cost circumstances for consultants led the evaluator to employ the higher cost estimates for consultants.

The supplies line-item estimate of $1,760 was included to cover paper and related materials for producing the updated evaluation plan, materials for each meeting, and materials for draft and final reports for the five contracted tasks.

The services line item of $27,444 assumed that on average, the project would expend $3,430 during each of eight months on such items as communication, photocopying, computer use, and postage. This rather large amount reflected the fact that the project would involve intensive collaboration and numerous teleconferences across a nationwide network of project personnel.

The total direct costs line item of $333,784 is the sum of the line items already enumerated, and the indirect costs figure was derived by multiplying $333,784 by the institution's indirect cost rate of 29 percent, which was a blended on-campus and off-campus rate. Adding the resulting $97,979 in indirect costs to the total direct costs yielded the bottom-line amount of $431,763.

In many evaluation contracts, budget details such as those just presented would be appended to the budget summary as budget notes for each cost item. In the military evaluation example, such notes were presented to the contractor's budget office but were not included in the budget submitted to the sponsor.

In the end, the evaluation team used about $50,000 less than the fixed-price amount of $431,763. This was largely due to the fact that the originally projected eight-month project became a six-month project. As mentioned in Chapter 19, the military organization took two months to process the contract, and the commandant's deadline for the final report remained fixed. Also, the evaluator no doubt somewhat overestimated the cost for the work.

When it became clear that the project would end with about a 13 percent surplus over what was expended, the evaluator informed the sponsor that his team would be willing to use the excess funds to perform additional services or return the unused funds. By this time, the evaluator had convinced the sponsor to allow site visits to bases other than those included in the original evaluation contract and noted that the original award had sufficient funds to cover the costs for that added work. In addition, the evaluator suggested that part of the excess funds be used later to support the field-testing and institutionalization of the new evaluation

system. The military organization's response was that it had no ability to accept a return of unused funds or authorize their use beyond the originally negotiated tasks. Ironically, the military organization issued an additional contract and associated award of $15,000 to support the evaluation team's added site visits. The evaluator's institution ended up with a windfall surplus of about $50,000. Such can be the nature of fixed-price evaluation agreements. But an institution can lose money on a fixed-price evaluation if the work has been underbid.

Other Types of Evaluation Budgets

Although the evaluation project just discussed was conducted under a fixed-price contract, it might have been pursued under some other type of agreement. In this section we define and discuss evaluation budgeting under grants, cost-reimbursable contracts, cost-plus agreements, cooperative agreements, and modular budgets.

Evaluation Budgeting Under Grants

An evaluation grant is a financial award to support a qualified evaluator to conduct a study that is of interest to the evaluator, contains societal value, lies within the sponsor's mission, and is seen to be at a fundable level. For example, a charitable foundation concerned with improving public schools was considering approving and funding an evaluator-initiated proposal to conduct a comparative evaluation of nineteenth- and twentieth-century school governance policies in a selected number of states. In this case the evaluator outlined goals and procedures for the desired investigation and sent his project plan and associated budget to the foundation for possible approval and funding. The sponsor saw the proposed study as worthy, related to its mission, competently planned and staffed, and financially supportable. Consequently, it awarded a grant to support the proposed evaluation. As with a fixed-price contract, the sponsor typically would expect the evaluator to use the funds wisely but not return unused funds.

Depending on the policies of the granting organization, the budget for a grant often may be quite general, as was the case with the budget submitted to the sponsor in the military evaluation example. The main difference between a grant and a fixed-price contract is related not to the funding, but to the sponsor's control over the study's tasks and reporting of findings. In the military evaluation example, the sponsor stipulated the tasks to be completed and required that findings be reported only to the USMC. Under a grant, the sponsor generally would not specify the study's tasks or control the release of findings. The granting organization's main interests would be to ensure that the proposed study has societal value and is related to the sponsor's mission, that the evaluator has the needed competence and record of professional responsibility, and that the requested amount of funds is available and appropriate to achieve the grant's objectives.

Another difference between a grant and a fixed-price contract concerns indirect costs. In the military evaluation example, the budget included a 29 percent indirect costs charge to cover such unspecified items as heat, light, custodial services, security, facilities, and fiscal accounting. Often granting organizations will pay little or no money for indirect costs. The rationale is that

the award is a charitable contribution to support the recipient's goals and program and is not a particular service to the granting organization. Furthermore, the position often is advanced that if the funded study is central to the work of the recipient organization, it should make an in-kind contribution, for example, by covering the associated overhead costs.

Budgeting under a grant has minimal risks to the sponsor and the evaluator. The sponsor typically requires only the level of accounting needed to ensure that funds were expended to achieve the project's approved purposes. If grant funds were used inappropriately, the granting organization could require that the funds be returned. The grantee's main risk in grant-related budgeting is that he or she might promise more than can be achieved with the granted amount of money. When this occurs, the grantee's efforts might fail and be judged negatively, or his or her organization might have to acquire additional funds internally or externally to get the job done.

Both sponsors and evaluators derive important benefits from securing grants for evaluations. Usually organizations that award such grants have substantial funds to expend on pursuing a clear and socially important mission. What they usually lack are the creative ideas, technical capabilities, field researchers, and supervisors to plan, carry out, and oversee important studies. Granting organizations meet these needs by attracting and issuing grants for high-quality, evaluator-initiated evaluation projects. These organizations also benefit because grants require minimal oversight of the grantee's operations and expenditures. Evaluators benefit from grants because they usually are flexible and leave much room for creativity and evolution in the supported project. Moreover, if the evaluator has inadvertently overestimated the needed funds despite careful planning, the evaluator's organization usually stands to retain the surplus.

Evaluation Budgeting Under Cost-Reimbursable Contracts

A cost-reimbursable evaluation contract is an agreement that the evaluator will account for, report, and be reimbursed for actual evaluation project expenditures. In the case of the military evaluation, under a cost-reimbursable contract, the evaluator would have billed the sponsor for only those funds actually expended. As long as the cumulative total did not exceed the original, agreed-on total funding amount and the evaluator had performed competently and responsibly, the sponsor would have paid the submitted bills. In the end, the sponsor would have kept any unused funds.

Typically a cost-reimbursable budget should break out cost estimates in considerably more specificity than was seen in the fixed-price military evaluation example. Also, the sponsor usually will require budget notes that explain each budget item. Reporting evaluation expenditures against a detailed line-item budget usually will provide a sufficient basis for reimbursement. A detailed work plan in the evaluation design provides a foundation for working out a detailed line-item budget and associated budget notes.

A cost-reimbursable budget is decidedly in the interest of funding organizations. Under such an arrangement, their risks are minimal. They pay only for completion of agreed-on work and do not have to pay any amount above the agreed-on ceiling price for the work. If the original budget proves insufficient to complete the project, the sponsor has the discretion to

award additional funds or terminate its support of the effort. Risks to the evaluator in this type of budget can be considerable, but they need not be. The main pitfall occurs when the work was underbid. In such a situation, the evaluator might have to incur the additional costs of completing the evaluation's scope of work. To the extent that the evaluator has made a valid estimate of costs or has obtained provisions for periodic updating of the budget, a cost-reimbursable arrangement usually provides a sound financial basis for conducting a defensible evaluation. Of course, under this type of budget, the evaluator's organization does not stand to reap a financial profit from the project.

Evaluation Budgeting Under Cost-Plus Agreements

A cost-plus agreement includes the funds needed to conduct an evaluation assignment plus an additional agreed-on charge for the evaluator's services outside the sphere of the contracted evaluation. Cost-plus budgets are of three types: cost plus a fee, cost plus a grant, and cost plus a profit.

Under a cost-plus-a-fee budget, the additional funds would be used to help sustain the contracting organization. Here the budget would specify an institutional sustainability fee, such as 1 percent of the direct costs. Under a cost-plus-a-grant budget, the additional funds could be used to support program functions—funding graduate students, research on evaluation, or an evaluation conference, for example. In preparing this part of the budget, the evaluator probably would provide separate line items for the project components that are outside the contracted evaluation tasks. The budget line item for such ancillary costs might simply be a specified charge or a line-item budget for the projected activities. Usually the sponsor would be asked to pay the requested cost-plus amount as a grant rather than as a cost-reimbursable charge. For-profit evaluation organizations typically employ cost-plus-a-profit budgets of some kind to reap financial gain from contracted evaluations. The profit margin might be hidden, as when it is incorporated into inflated personnel hourly charges or overhead. Or it might be explicitly included as a percentage of direct costs. In either case, we advise the contracting organization to be up front in disclosing how it budgeted for its profit margin.

A cost-plus budget can be built into a grant, a contract, or a cooperative agreement. In the last case, the agreement might call for the evaluator to charge only the actual costs of his or her part in the evaluation plus the agreed-on added amount for his or her organization. We see a cost-plus budget as ethical and appropriate as long as the evaluator and sponsor explicitly settle on such a provision in the original agreement. We discourage evaluators from surreptitiously inflating their budgets to reap a sizable surplus or fund unauthorized activities.

Neither the sponsor nor the evaluator incurs remarkable risks using cost-plus budgets as long as there is appropriate disclosure of the basis for the funding request. By buying into such an arrangement, the sponsor will purchase the desired evaluation work and willingly provide the contracting organization with additional funds for its other uses. The contracting organization will get the funds it needs to carry out the evaluation assignment plus additional funding to further its organizational viability and accomplishments. The risks and benefits related to the cost of the evaluation portion of the budget are the same as those defined

earlier, depending on whether that part of the budget is a grant, a fixed-price agreement, or a cost-reimbursable agreement.

Evaluation Budgeting Under Cooperative Agreements

A cooperative evaluation agreement is an arrangement for the evaluator and sponsor to collaborate in conducting the evaluation. Under such an agreement, the evaluator and sponsor share responsibility for and authority over carrying out the evaluation work. They also share the resources needed to discharge their joint and individual responsibilities.

Cooperative agreements can be beneficial when there is appropriate differentiation of sponsor and evaluator roles. For example, the sponsor can facilitate the evaluator's work by such contributions as providing office space, clerical support, assistance in gathering data, and support for stakeholder involvement and cooperation. In addition, the collaboration that occurs in a cooperative agreement–based evaluation may promote the sponsor partner's ownership and use of evaluation findings.

However, cooperative agreements also have the potential to thwart the conduct of an effective and professionally defensible evaluation. The basic issue is that such agreements set aside the evaluator's independence. Unless evaluator and sponsor roles are differentiated and delineated clearly, appropriately, and in enforceable ways, the evaluator can be put in the unfortunate position of having responsibility for the study's success and integrity but lacking authority to do all that is needed. In such instances, the sponsor may be positioned to act on its conflicts of interest. For example, it might censor, inappropriately edit, or withhold release of embarrassing reports. It might insist on hiring persons for the evaluation team who are not the most qualified for the work. Or it might not discharge its responsibilities in a timely fashion. Also, the sponsor might inappropriately control the use of evaluation funds. As with empowerment evaluations, an evaluation under a cooperative agreement can place the evaluator in the position of lending technical assistance and undeserved credibility to a sponsor who is in the evaluation's driver's seat. In general, a cooperative evaluation agreement is a threat to the evaluator's ability to carry out and report on an independent study.

Of course, part of the cooperative agreement should focus on the budget. Often the allocated funds will be divided between the evaluator and the sponsor. The evaluator should delineate his or her funding requirements and stipulate in the contract that he or she will have appropriate control over use of the necessary funds. The agreement should also be clear about funds and related resources that the sponsor will expend in carrying out its part of the evaluation work.

The sponsor benefits from a cooperative agreement by being in a position to strongly influence the evaluation. The evaluator especially benefits when the sponsoring organization facilitates the evaluation, comes to value its findings, and uses the findings for decision making and accountability purposes.

Nevertheless, we think cooperative agreements are the weakest and most problematic type of evaluation arrangement. Considering the potential for problems in such arrangements, it is difficult to include a sufficient set of safeguards. If an evaluator must enter into a cooperative

agreement, it is essential to negotiate a clear and appropriate contract according to the guidelines set out in Chapter 21. The agreement should stipulate that the evaluator will have appropriate authority over expending necessary funds. Also, it is a good idea to secure an agreement that an independent metaevaluator will oversee and report on the effort.

Evaluation Budgeting Under Modular Budgets

Modular evaluation budgets delineate the required funding for each part of a designed evaluation project or for each project year or other time period. It can be important to modularize an evaluation budget for three main reasons: (1) a sponsor might be uncertain about how much of a proposed evaluation it can or would want to fund, (2) several prospective sponsors might want to share the funding of the evaluation work, or (3) the evaluator might be able to be explicit about certain modules of the project but only tentative about others. In such cases, there is a need to provide a budget for each task of the designed evaluation project or for each year or other period of work. A modular presentation is appropriate with all the other budget types discussed in this chapter.

We can see how modular budgets work in evaluations that have several tasks by revisiting the military evaluation example. That evaluation had five main tasks. Because the sponsor chose to fund all five tasks, it required only a bottom-line amount for the total job. To arrive at this amount, the evaluator developed and submitted the line-item budget shown in Table 20.1. Had the sponsor required a modular budget, the evaluator could have constructed one based on the evaluation design's delineation of evaluation activities. Table 20.2 is an illustrative framework

Table 20.2 Illustrative Framework for Constructing a Modular Evaluation Budget Showing Line Items and Tasks

Line Item	Task 1	Task 2	Task 3	Task 4	Task 5	Total
A. Personnel						
1. Salaries						
2. Fringe benefits						
3. Total personnel						
B. Travel						
C. Consultants						
4. Honoraria						
5. Travel plus support services for consultant team leaders						
6. Total consultants						
D. Supplies						
E. Services (telephone, photocopying, and postage)						
F. Total direct costs						
G. Total indirect costs						
H. Total project costs						

Table 20.3 Illustrative Framework for Constructing a Modular Evaluation Budget Showing Line Items and Years

Line Item	Year 1	Year 2	Year 3	Year 4	Year 5	Total
A. Personnel						
1. Salaries						
2. Fringe benefits						
3. Total personnel						
B. Travel						
C. Consultants						
4. Honoraria						
5. Travel plus support services for consultant team leaders						
6. Total consultants						
D. Supplies						
E. Services (telephone, photocopying, and postage)						
F. Total direct costs						
G. Total indirect costs						
H. Total project costs						

Table 20.4 Illustrative Framework for Constructing a Modular Evaluation Budget Summarizing Costs by Task and Year

Task	Year 1	Year 2	Year 3	Year 4	Year 5	Total
Task 1						
Task 2						
Task 3						
Task 4						
Task 5						
Total						

for constructing a modular budget that breaks out evaluation tasks. Use of such a framework generates line-item costs for each project task and the total project. Table 20.3 shows how a budget would be broken out by line item and year (or other time period). Table 20.4 shows how a budget could be summarized by task and year (or other time period).

Summary of Budget Types

Table 20.5 summarizes the budget types discussed in this chapter. Often evaluators will not have a choice of the type of budget to use because the dispenser of funds usually dictates the type of budget it will accept. However, there are occasions when the evaluator can propose a type of budget to be followed. In these situations, we suggest that the evaluator carefully

Table 20.5 Summary of Budget Types

Budget Type	Key Points	Risks to the Sponsor	Risks to the Evaluator	Benefits for the Sponsor	Benefits for the Evaluator
Fixed-price contract	A firm amount is to be paid to the evaluator for performing the sponsor's defined scope of work.	The sponsor might pay much more than the actual cost if the needed funds were overestimated.	The evaluator might incur a financial loss if the needed funds were underestimated.	The sponsor is likely to obtain the needed services at an acceptable price.	The evaluator might gain a profit if the needed funds were overestimated.
Grant	A grant is awarded for an approved study with minimal oversight and control by the sponsor and often without reimbursement for indirect costs.	Risks are minimal if proper accounting ensures proper expenditure of funds.	Risks are minimal if the evaluator receives sufficient internal and external resources to meet study objectives.	This type of budget means that a sponsor can support projects with relevance to the sponsor's mission while supplying minimal oversight.	There is a flexible source of funds that allows the evaluator to creatively pursue the study objectives.
Cost-reimbursable contract	Payment is restricted to actual expenditures up to a given limit.	Risks are minimal because payments are for successful completion of tasks up to a specified limit.	Success is jeopardized if the evaluator underbid the job; otherwise, risks are minimal.	Use of this budget type entails funding achievement of the sponsor's objectives at cost and not above a limit.	This budget type usually provides a sound financial basis for completing a defensible study.
Cost-plus agreement	Funds cover the evaluation's needs plus a portion to support other contractor functions.	Risks are minimal, because the funds will purchase the desired study and cover an additionally approved award to the contractor.	Risks are minimal as long as the evaluator accurately estimated costs for the evaluation and properly disclosed the additional charge.	The sponsor will purchase the desired evaluation and knowingly provide the evaluation contractor with additional financial assistance.	The evaluator receives the funds needed to conduct the evaluation plus additional money to support organizational sustainability and accomplishments.
Cooperative agreement	Contractor and sponsor share responsibility and authority in conducting the evaluation.	Risks are moderate. Although the sponsor retains much control over program and financial decisions, there is potential for conflict with the contractor concerning responsibility and authority.	Risks are substantial because the evaluator loses independence and the sponsor can exert strong, even inappropriate influence over the work.	The sponsor can strongly influence the contractor's decisions and actions to ensure that sponsor interests and needs are fully served.	The sponsor might perform important evaluation tasks related, for example, to stakeholder cooperation, clerical support, and data collection.
Modular budget	The budget is broken out by major task, project year or other term, or both.	There are none, because cost estimates are delineated and better justified.	Risks are moderate. Although the breakout strengthens mutual understanding of the budget, it also might tempt the sponsor to drop certain important parts of the evaluation.	This type of budgeting helps the sponsor decide what components to fund or whether to delay decisions about some of them.	This approach provides a basis for allocating funds to tasks, time periods, or both; for giving general estimates for some of the work; and for delaying firm estimates until later.

consider what type of budget would best serve the purposes of the evaluation and the interests of the involved parties. Key considerations are risks and benefits to the sponsor and contractor. In general, we advise against employing a cooperative agreement. Under appropriate safeguards, however, it can be advantageous for the evaluator to secure the sponsor's assistance in involving stakeholders, housing evaluation operations, providing access to pertinent files, clearing the way for data collection, promoting use of findings, and so forth. The key pitfalls to avoid in cooperative agreements are compromising the evaluator's independence and denying the evaluator the authority required to fulfill his or her evaluation responsibilities. No matter what type of budget is employed, it can be appropriate to break it out not only by line item but also by task and year or other work period.

Generic Checklist for Developing Evaluation Budgets

We conclude this chapter by presenting and discussing the generic checklist for budgeting evaluations displayed in Exhibit 20.1. The checklist is intended for use in both constructing and reviewing a budget and proposed set of financial agreements. This checklist lists ten major tasks to carry out in establishing a sound financial basis for a projected evaluation. Each task is divided into specific items to consider during the budget development process. In general, the checklist is best applied by considering the ten tasks in the given order. However, one often will skip and later return to some of the specific items. We advise users of the checklist to cycle through it repeatedly during the budget development process. We provide commentary following the exhibit on each of the checklist's ten tasks.

Exhibit 20.1 GENERIC CHECKLIST FOR DEVELOPING EVALUATION BUDGETS

1. Ensure that the evaluation design includes sufficient detail for building a sound budget. (Check all that apply.)
   - _____ Tasks
   - _____ Activities
   - _____ Personnel and consultants
   - _____ Nonpersonnel resources
   - _____ Funding period and schedule
   - _____ Subcontracts
   - _____ Provisions for updating the budget as appropriate
2. Determine the appropriate type(s) of budget agreement. (Check all that apply.)
   - _____ Grant
   - _____ Fixed-price contract
   - _____ Cost-reimbursable contract

_____ Cost-plus-a-fee agreement
_____ Cost-plus-a-grant agreement
_____ Cost-plus-a-profit agreement
_____ Cooperative agreement
_____ Modular budget

3. Determine the required level of budget detail. (Check all that apply.)

_____ Line-item budget
_____ Line items by task
_____ Line items by year (or other work period)
_____ Tasks by year (or other work period)
_____ Total budget only
_____ Breakout of the local contribution
_____ Budget notes

4. Determine pertinent cost factors. (Check all that apply.)

_____ Budget ceiling
_____ Allowance for pre-award costs
_____ Hiring costs
_____ Name or job title and daily salary rate for each staff member
_____ Name or job title and hourly salary rate for each staff member
_____ Fringe rates for each category of staff
_____ Number of workdays for each staff member
_____ Number of work hours for each staff member
_____ Daily rate for staff per diem
_____ Projected number of staff trips
_____ Projected average travel cost per staff trip
_____ Name or job title and daily rate for each consultant
_____ Number of workdays per consultant
_____ Name or job title and hourly rate for each consultant
_____ Number of work hours for each consultant
_____ Projected number of consultant trips
_____ Projected total travel days for consultants
_____ Daily per diem rate for the consultants
_____ Projected average travel cost per consultant trip
_____ Indirect cost rate
_____ Factor for annual staff salary increments
_____ Factor for the annual level of inflation
_____ Institutional sustainability fee factor
_____ Profit factor

_____ Other

5. Determine line items. (Check all that apply.)

_____ Personnel salaries
_____ Personnel fringe benefits
_____ Total personnel
_____ Travel
_____ Consultant honoraria
_____ Consultant travel
_____ Consultant materials and other support
_____ Total consultant costs
_____ Supplies
_____ Telephone
_____ Photocopying and printing
_____ Computers
_____ Postage
_____ Total direct costs
_____ Total indirect costs
_____ Institutional sustainability fee
_____ Supplemental grant
_____ Contractor profit
_____ Subcontracts
_____ Other costs

6. Group line items for convenience. (Check all that apply.)

_____ Personnel
_____ Travel
_____ Consultants
_____ Supplies
_____ Services
_____ Subcontracts
_____ Total direct costs
_____ Total indirect costs
_____ Total project costs
_____ Budget notes

7. Determine the local contribution, if any. (Check all that apply.)

_____ Reduction or elimination of indirect cost charges
_____ Contributed time of staff members
_____ Institutional funding of certain direct expenses
_____ Other

8. Compute costs and charges. (Check all that apply.)
   - _____ Charges by year (or other work period)
   - _____ Charges by project tasks
   - _____ Charges by subcontract
   - _____ Overall charges
   - _____ Local contribution
   - _____ Budget notes to be added
   - _____ Independent budget review to be obtained
9. Provide for institutional fiscal accountability. (Check all that apply.)
   - _____ Responsibility for internal accounting
   - _____ Responsibility for financial reporting
   - _____ Provision for internal auditing of project finances
10. Clarify requirements for payment. (Check all that apply)
   - _____ Funding source and contact persons
   - _____ Financial reporting requirements
   - _____ Schedule of financial reports
   - _____ Amounts and schedule of payments

Task 1: Ensure That the Evaluation Design Is Sufficiently Detailed

A sound evaluation design is a precondition for developing a functional, complete, and defensible evaluation budget. When starting the budget development process, the evaluator should ensure that the evaluation design is as fully developed as the situation warrants. The checklist sets out the essential design components: the major tasks, activities, staff and consultants, nonpersonnel resources, the funding period and schedule, and any subcontracts. Such design elements should be clearly defined, so the evaluator can confidently assign costs for carrying out the evaluation. If any essential design elements are missing or unclear, the evaluator should, as feasible, improve the design as needed. In the case of a responsive or formative evaluation in which all details cannot be specified in advance, the evaluator should seek agreement from the sponsor that the evaluation budget will be updated periodically as the evaluation design and activities evolve.

Task 2: Determine the Type of Budget Agreement

There are several types of budget agreements, each entailing different costing approaches and levels of detail. The evaluator should clearly determine the type of budget to be employed. If the budget is subject to periodic updates, as in the case of a responsive evaluation, the evaluator should have the sponsor confirm in writing that payments will be made in accordance with the periodic budget updates.

Task 3: Determine the Needed Level of Budget Detail

The evaluator should next determine how much budget detail to provide. Partly this will depend on the type of budget being employed. For example, the sponsor of a grant or fixed-price contract may require only a total cost figure or a general breakout of estimated costs and charges. Other sponsors and the evaluator's home institution typically require much more detail. Task 3 shows that the more detailed presentations include a detailed line-item budget. It can be appropriate to divide any type of budget by task, year (or other work period), and/or local and sponsor contributions. Even when there are no requirements for delineating costs, the evaluator often can benefit by preparing and using the budget breakouts as management tools. Whether or not the sponsor requires a detailed budget, the evaluator typically should prepare notes that explain each budget item. Normally the evaluator's budget office will require such detailed budget information.

Task 4: Determine Pertinent Cost Factors

Task 4 contains an extensive list of potential cost factors. Evaluators should carefully consider the full range of these factors and possibly others. Doing so helps ensure that they will have at hand the rates needed to compute costs for the full range of projected cost items.

The first item in this task concerns a possible budget ceiling and is especially noteworthy. We advise evaluators to investigate whether the sponsor has in mind a limit for the amount of funding for the evaluation and, if so, to at least attempt to determine what it is. Often the sponsor will have established such a limit. Sometimes the limit will be published. In other cases, the sponsor might reveal it if asked. Information on an evaluation's funding limit can be very useful in building an evaluation budget. It helps the evaluator consider whether the evaluation is feasible given funding restrictions. It also helps the evaluator determine how much of a local contribution might be necessary to supplement the sponsor's funds to conduct a sound evaluation. For example, the evaluator's organization might decide to reduce its indirect cost rate so that the evaluator will be able to use nearly all of the funds to be provided by the sponsor.

Not all factors in this task need to be or should be incorporated into an evaluation budget. For example, the budget should be based on either daily rates or hourly rates for staff and consultants—not both. Employing both of these in a given budget is likely to confuse the sponsor and other reviewers. Moreover, some evaluations might not involve travel or employment of consultants, and many will not include an institutional sustainability fee or a profit factor. In addition, only multiyear projects need to include increases to cover inflation and salary increments.

Notice that there are separate line items for estimated travel costs for staff and consultants. This reflects the possibility that travel cost rates for staff may differ substantially from those for consultants. The evaluator may be able to bargain for lower travel rates for staff because of the projected high volume of their travel. Or consultant travel costs could be lower than staff travel costs if the consultants live close to the evaluation site and staff are located further away.

Staff and consultant pay rates in this task assume that those who will carry out the evaluation are identified in the evaluation design. When this is the case, it is possible to specify exact pay rates. Some evaluation designs, however, may identify only the job titles of personnel to be involved, with some staff selections to occur following funding. In these instances the evaluator should identify and provide cost estimates for each staff position.

Table 20.6 is a worksheet for estimating personnel costs by category. The daily rates are only examples, because these should be determined according to circumstances. Separate estimates are given for core staff, graduate students (assuming that the evaluation is university associated), and consultants. One reason for this is that universities do not charge project sponsors for or give fringe benefits to graduate students and consultants. Another reason for separating out consultants is that government authorities in the United States do not require organizations to pay unemployment compensation taxes for the contracting organization's consultants if they are independent contractors and not regular staff. By showing clearly in the budget that the consultants are independent contractors and not members of the evaluation project's salaried staff, the writer of the budget provides his or her parent organization with a measure of protection from federal auditors who might later want to cite the organization for not paying into a government unemployment compensation fund related to the amounts of money that the parent organization paid to the consultants.

Table 20.6 Worksheet for Determining Costs for Categories of Personnel

	Task 1		Task 2		Task 3		Total	
Personnel Categories	**Days**	**Cost**	**Days**	**Cost**	**Days**	**Cost**	**Days**	**Cost**
Core evaluation staff								
Principal investigator: $800 per day								
High-level methodologists: $600 per day								
Field researchers: $300 per day								
Technical support staff: $250 per day								
Clerical staff: $100 per day								
Total core staff without fringe								
Fringe rate: 40%								
Total senior staff: fringe loaded								
Graduate students								
Advanced students: $175 per day								
Entry-level students $125 per day								
Total graduate students								
Consultants								
High-level consultants: $800 per day								
Medium-level consultants: $400 per day								
Total consultants								
Total personnel								

In evaluations that span multiple years, the evaluator can complete a useful separate worksheet for each project year. In these cases, the evaluator should project and provide for annual increments related to inflation and salary increases. Depending on recent economic trends, such increments might range between 0 percent and 5 percent.

Another item in this task that merits special mention concerns the basis and rate for indirect costs. Most contracting organizations have an established indirect cost rate for inclusion in proposals, and usually sponsors agree to pay that rate. However, some sponsors agree to pay indirect costs on only part of the direct costs (or not at all). For example, the sponsor might decide to pay travel costs directly rather than have them included in the evaluation budget. This arrangement has the effect of lowering the indirect cost charges to the sponsor. In such a case, the evaluator would provide the sponsor with an estimate of travel costs, but not include the estimate in the evaluation budget. By agreement, evaluation team members would submit travel bills to the sponsor, who would then reimburse them. In general, the evaluator should be alert to the prospective sponsor's policies and practices concerning direct and indirect costs.

Task 5: Determine Line Items

Task 5 is to determine all items to be covered by the evaluation budget. We have included items that commonly appear in evaluation budgets. We suggest that evaluators check all items that apply to their particular evaluations, and subsequently consider whether other items should be added to the list. Task 5 is highly important in the budget development process. Its purpose is to ensure that evaluators do not fail to project all of an evaluation's costs.

Task 6: Group Line Items for Convenience

In task 6 the evaluator groups the items identified in task 5 into typical budget categories. The intent is to provide the sponsor with an efficient presentation of budget information. The items in task 6 are fairly standard in evaluation budgets, but they can be modified according to the preferences of a sponsor and evaluator.

Task 7: Determine the Local Contribution, If Any

Task 7 advises the evaluator to consider whether her or his parent organization should make a contribution to the evaluation. In certain agreements, such as grants, the sponsor may require a contribution from the contractor, such as elimination or reduction of indirect cost charges. Or the evaluator might know the approximate limit of available funds from the sponsor and discern that charging the full indirect cost rate would preclude receiving funds needed for a fully responsive and technically sound evaluation. In such cases, the contractor might agree to lower or waive its indirect cost rate. Other contractor contributions might be in sharing the cost of personnel and directly funding certain evaluation tasks. Whatever the contractor's local contribution, it is wise to report it along with the submitted budget. Accordingly, the evaluator should consider including a "Local Contribution" column heading in each project year's line-item budget.

Task 8: Compute the Evaluation's Costs and Charges

Provided that the preceding tasks have been accomplished, the evaluator can accurately compute the evaluation's costs and charges. The evaluation design, sponsor's needs, and requirements of the evaluator's home institution will determine the nature of the required budget displays. A basic rule is to develop the budget first at its most detailed level, then aggregate as appropriate. For example, if the evaluation has multiple tasks and will be conducted across multiple years, the evaluator should begin by developing a line-item budget, broken out by tasks for each year. At this stage, the evaluator should consider padding certain budget items, to a modest level, to make sure the evaluation is not underbid. Each year's budget display should be backed up with an appropriate set of explanatory budget notes. The master set of figures can subsequently be aggregated in various ways to serve the interests of different audiences. Also, a summary set of budget notes can be prepared and appended to the evaluation proposal. When the budget charts and notes have been completed, the evaluator should obtain critiques of the budget and use them to correct and finalize the budget.

Task 9: Provide for Institutional Fiscal Accountability

In an evaluation proposal, it is appropriate to provide the sponsor with assurances that the evaluation's finances will be appropriately and professionally monitored, controlled, and reported. Even in small, single-evaluator studies, the evaluator will need to keep track of expenditures and carry out some type of internal accounting. In sizable evaluations that are conducted through universities or other contracting organizations, the evaluator should inform the sponsor of the agents who will control and conduct internal audits of the use of funds. In addition, the evaluator needs to inform the sponsor of the office and employees that will be making financial reports and answering the sponsor's queries. All such arrangements are in the interest of effecting sound fiscal accountability and a professional relationship with the sponsor. In many cases, the sponsor will explicitly require the submission of such information as part of the evaluation proposal. Even if the sponsor makes no such requirement, the evaluator is wise to provide for and report plans for maintaining fiscal accountability.

Task 10: Clarify Requirements for Payment

As a final budget preparation task, the evaluator should reach an agreement with the sponsor on sponsor and contractor responsibilities in regard to payment for the evaluation. Coming to such an agreement, particularly in the case of a large or longitudinal study, may require the evaluator and sponsor to seek legal advice before a contract is finalized. Moreover, the evaluator needs to identify the office and contact persons in charge of making the payments on behalf of the sponsor. The evaluator will want to clarify the sponsor's requirements in regard to financial reports and when they should be submitted. Also, the evaluator should clarify the amounts and schedule of payments to be received from the sponsor. Although it is not done often, the evaluator is advised to summarize, in a budget note, her or his understanding about the matters just discussed.

This concludes our discussion of Exhibit 20.1. We have shown that budgeting for evaluation studies is a complex process, because different types of budgets are involved in serving a wide range of different evaluation assignments. Nevertheless, one can follow a systematic, step-by-step process to arrive at the needed budget. The Generic Checklist for Developing Evaluation Budgets is recommended as a tool for making the budgeting process efficient, technically sound, and ethical.

Summary

In this chapter we undertook to present a wide-ranging yet practical discussion of budgeting for evaluations. We stressed that budgeting should be a fully ethical process and discussed some of the ethical pitfalls in evaluation budgeting. We referred to Chapter 19's military evaluation case to provide an example of a fixed-price, line-item budget. Subsequently, we defined budgets keyed to grants, contracts, and cooperative agreements. In addition, we discussed different types of budgets, including fixed-price, cost-reimbursable, and cost-plus agreements. We also noted that any of these types of budgets can and often should be broken out by year (or other work period) and task. Finally, we presented and explained a checklist for use in systematically developing evaluation budgets.

Budgeting is a key part of evaluation work, and it should be done ethically and well. Additional valuable sources of information on evaluation budgeting are *The Contract and Fee-Setting Guide for Consultants and Professionals* (Shenson, 1990) and *A Checklist for Developing and Evaluating Evaluation Budgets* (Horn, 2001).

REVIEW QUESTIONS

1. Why is it possible to give a budget that is essentially fully specified for a preordinate evaluation, but not for a responsive evaluation?
2. There are connections between evaluation design and budgetary considerations. Referring to Chapter 19, state at least three such connections, together with the part that evaluator-sponsor collaboration might play in strengthening budgetary decisions or even evaluation design decisions.
3. What are the essential ethical imperatives for developing evaluation budgets?
4. What are at least four unethical evaluation budgeting practices that should be steadfastly avoided?
5. What are the attractions of a fixed-price agreement for both evaluator and sponsor?
6. List, with a brief comment on each, five key decisions that need to be made in the course of developing an evaluation budget.
7. List at least ten cost factors to be determined in the budget development process, and explain the importance of each factor in completing the budget.

8. What do you understand by a cost-reimbursable evaluation contract? What do you understand by a cost-plus agreement?
9. On the one hand, why is independence a serious issue with cooperative agreements? On the other hand, what does an evaluator stand to gain from a cooperative evaluation agreement?
10. What are modular evaluation budgets, and under what circumstances can they prove useful?

Group Exercises

Exercise 1

"Cooperative agreements are the weakest and most problematic type of evaluation arrangement." Discuss this statement from the budgetary point of view, and reach conclusions about its validity (or otherwise).

Exercise 2

In this chapter we have provided a checklist of ten tasks that we advise evaluators to draw on during any budget development process. As a group, reread this section of the chapter, memorizing salient features of each of the ten tasks. Then appoint a leader whose job it is to ask a group member both to describe in some detail the reasons for conducting a particular task and to comment on its importance from the viewpoint of both evaluator and sponsor. After all ten tasks have been tackled, the group leader will open the discussion to encompass any other matters pertaining to budgeting for evaluations.

Exercise 3

Ask a member of your group to obtain and brief the other members on a proposal for an evaluation, including its budget and associated budget notes. Then, as a group, analyze the budget in terms of the proposal's type of evaluation agreement, the type of evaluation budget, and the presence or absence of ethical safeguards. Finally, employ the checklist for evaluation budgeting in Exhibit 20.1 to characterize and assess the adequacy of the proposal's evaluation budget.

Note

1. In actuality, midway in the evaluation the sponsor reversed its previous decision not to allow the evaluator to interview marines away from USMC headquarters and accordingly issued a supplementary award of approximately $15,000 to support the conduct of such interviews.

Suggested Supplemental Readings

Cronbach, L. J., & Associates. (1980). *Toward reform of program evaluation*. San Francisco, CA: Jossey-Bass.

Horn, J. (2001). *A checklist for developing and evaluating evaluation budgets*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/evaluationbudgets.pdf

Shenson, H. L. (1990). *The contract and fee-setting guide for consultants and professionals*. Hoboken, NJ: Wiley.

CHAPTER 21

CONTRACTING EVALUATIONS

This chapter is intended to assist evaluators and their clients with developing and applying sound, enforceable agreements for successfully executing, reporting on, and following up on evaluations.

Closely tied to designing an evaluation (Chapter 19) and evaluation budgeting (Chapter 20) is evaluation contracting. Although we have presented these chapters as somewhat separate entities, in fact aspects of them are closely interrelated. Imagine that an evaluation design has been agreed to by both evaluator and client and also that budgetary arrangements have been accepted. There now remains the significant step of negotiating a contract for proceeding that is satisfactory to both parties. Such a contract must provide ample evidence that the evaluation will be conducted usefully, professionally, ethically, and legally based on mutual trust and concrete advance agreements between evaluator and client.

Negotiating an evaluation contract or memorandum of agreement is one of the most important steps to ensure an evaluation's success. This process of evaluation contracting establishes a trusting relationship between an evaluator and a client and grounds their agreements in a written contract or memorandum of agreement. Such instrumentalities are vitally important for documenting essential agreements; holding both parties accountable for discharging their agreed-on responsibilities; and resolving disputes that may emerge concerning management, funding, implementation, reporting, or a host of matters. In this chapter we define terms associated with advance evaluation agreements, explain the role and importance of such agreements, discuss the process of negotiating agreements, and present a checklist for developing or assessing evaluation agreements.

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **Definitions of evaluation contracts and memorandums of agreement**
- **Core requirements of evaluation contracts and memorandums of agreement**
- **The rationale for negotiating advance agreements to guide and govern evaluations**
- **The process of negotiating evaluation agreements, including stakeholder engagement**
- **Advice for addressing organizational contracting requirements**
- **Political issues in evaluation contracting**
- **The detailed contents and uses of a checklist for evaluation contracting**

Definitions of Evaluation Contracts and Memorandums of Agreement

An evaluation agreement may take the form of a formal contract or a less formal memorandum of agreement. Both forms of agreement should provide a framework of mutual understanding for proceeding with evaluation work. The formal contract is more applicable in external evaluations, and memorandums of agreement are better suited to internal evaluations. Preferably, both should be printed.

Such evaluation agreements are between two parties. Typically these are a client—who needs an evaluation of his or her program and funds an evaluator to conduct the evaluation—and a selected evaluator. Other evaluation agreements may be between an evaluator and a sponsor who will fund the evaluation, so that a designated client group can receive and use the evaluation's findings. Although evaluation contracts should, as feasible, reflect inputs from the full range of stakeholders, none of them can be an independent third-party signer of the formal evaluation contract. A contract among three or more parties would create confusion—especially in regard to appropriate allocations of authority and responsibility—and would be hard if not impossible to enforce. The evaluator may and often should, however, include in the formal contract commitments concerning the involvement of stakeholders, such as a guarantee of confidentiality in regard to their completion of questionnaires or participation in interviews or a promise to provide them with a summary of the evaluation's findings. It would be the evaluator's responsibility to see to the enforcement of such commitments to stakeholders.

Evaluation Contracts

We define an evaluation contract as a legally enforceable, written agreement between an evaluator and a client or sponsor concerning an evaluation's specifications and both parties' responsibilities. An evaluation agreement could be only oral and still be legally enforceable, but for practical reasons, we recommend that a contract be written. Written contracts provide concrete information for review. They help reduce later possibilities for misunderstandings and disputes far better than do oral agreements, which each party may remember differently. It is desirable that a client and evaluator define, at least in general terms, what would constitute a breach of contract by either party and what consequent actions might be taken. A contract should be consistent with pertinent federal and state laws and local regulations. Also, it should stipulate bases and procedures for canceling or amending the contract prior to, during, or following an evaluation. Before finalizing a contract, it is often appropriate to have it reviewed by an attorney.

Evaluation Memorandums of Agreement

A memorandum of agreement is similar to an evaluation contract except that the former is less formal. It is an evaluator's write-up of her or his understanding of agreements reached with a client for proceeding with an evaluation. At a minimum, it should denote what is to be done, by whom, how, where, when (including a completion date), and with the support of what funds

and other needed resources. The memorandum could even be drawn from the minutes of an evaluation planning meeting.

An evaluator writes the memorandum of agreement and submits it to a potential client, typically an official of the evaluator's organization. The client then reviews and approves, amends, or rejects the proposed agreement. Such memorandums are especially applicable in cases of internal evaluations where a formal written contract would be atypical and awkward. For example, having met with a school district superintendent, and having discussed a particular evaluation assignment, a district's evaluator would then prepare a memorandum reporting their agreements and submit it for the superintendent's approval or amendment. Beyond stating the agreements reached in the meeting, such a memorandum often would set out the proposed evaluation's details. The evaluator would ask the superintendent to approve the contents of the memorandum in writing or engage in further deliberation.

Core Requirements of Evaluation Contracts and Memorandums of Agreement

According to the Joint Committee on Standards for Educational Evaluation (1994), evaluators should write evaluation agreements that contain "mutual understandings of the specified expectations and responsibilities of both the client and the evaluator" (p. 87). The Joint Committee continued, "Having entered into such an agreement, both parties have an obligation to carry it out in a forthright manner or to renegotiate it. Neither party is obligated to honor decisions made unilaterally by the other" (p. 87). It would be a mistake for a client or evaluator to act unilaterally in any matter for which the evaluation agreement requires joint decision making. Moreover, neither the client nor the evaluator should materially change the evaluation's design, scope, timeline, or budget without reaching an appropriate agreement with the second party to amend the contract. It is essential that an agreement be keyed to the evaluation assignment, design, and budget, as well as to professional standards for sound evaluations. It should be as comprehensive as past evaluation planning allows. It should be stated clearly, recorded in writing, and signed by both parties. Usually it is prudent to cite the evaluation design, the budget, and selected professional evaluation standards as clear parts of the agreement.

An evaluation agreement should be negotiated and completed before starting an evaluation study. It should cover as completely as possible the full range of issues that might impede an evaluation or cause it to fail. Often many of these issues are not accounted for in the evaluation design, and it would be a mistake to consider the evaluation design to be a complete written agreement. Among matters for agreement are the evaluation design, data collection, and reporting schedule; access to needed information; protection of evaluation participants; individual and joint responsibilities for conducting the evaluation; security of the obtained information; evaluation reports and other deliverables; right-to-know audiences; agreements by certain stakeholders to cooperate with the evaluation; editorial responsibility and authority; dissemination of reports; arrangements to foster use of findings; funding; uses of the evaluation for educational purposes; and publication of evaluation results or other

publishable features of the evaluation. Other agreements should define the standards for judging the evaluation, the study's objectives and scope of work, safeguards against the possible corruption of the evaluation, deliverables and their due dates, protocols to be observed in collecting and reporting information, provisions for keeping financial records and reporting financial information, and the terms of compensation for the work. Both evaluator and client will have important responsibilities for achieving a professionally defensible and effective evaluation. These responsibilities should be clearly defined and differentiated. Moreover, areas of authority for each party should be defined pursuant to the party's responsibilities.

Related to the preceding discussion, neither a client nor an evaluator should act unilaterally in matters that should have been but were not addressed in the evaluation agreement. Some examples could include providing particular program stakeholders with interim findings while withholding them from other segments of the evaluation's audience; editing a report, prior to its release, to exclude potentially embarrassing or counterproductive findings; releasing preliminary findings to the press; and delaying indefinitely the release of the final report. In general, evaluators and clients should comport themselves professionally, and should candidly communicate with each other regarding all matters concerned with conducting, reporting, and using results of a contracted evaluation—including those that were not anticipated when the contract was negotiated.

Provisions for Modifying Agreements

Although written evaluation agreements should be as explicit as possible, they also should allow for appropriate, mutually agreeable adjustments during the planning and execution of an evaluation. Such agreements will be more tentative in formative evaluations than they will be in summative ones. However, even in a tightly designed, preordinate evaluation, it would be a mistake to make the contract so detailed that it impedes the evaluator's creativity. Also, an evaluator and client should bear in mind the study's purposes when applying an evaluation agreement. They should not adhere so rigidly to a contract that they cannot make—or are unduly delayed in making—needed changes.

Rationale for Evaluation Contracting

Evaluators need to be skillful in negotiating advance written evaluation agreements for a number of reasons. In general, such agreements clarify understandings, build rapport and trust in the process of reaching agreements, help prevent disputes between clients and evaluators, and provide a basis for resolving any future evaluation-related disagreements. Advance agreements can mean the difference between an evaluation's success and failure. They can help reduce and resolve a wide range of possible day-to-day misunderstandings or lapses in memory, or other sources of potential conflict. Without advance agreements, the evaluation process is subject to misunderstandings, disputes, efforts to compromise findings, attacks, or even a client's withdrawal of cooperation and funds. In one high-stakes study, reference to the advance agreements on editorial authority and release of findings helped prevent the client

from burying the report or rewriting it. It helped the evaluators give assurance to members of the evaluation audience that the study had provided for and maintained its independence and objectivity. Clients may also reference sound contracts to convince their policy board or constituents that their institution contracted for sound, clearly defined evaluation services and can hold the evaluators to the agreements.

Stakeholder Engagement in Evaluation Contracting

We stress the importance of consulting with stakeholders prior to finalizing an evaluation agreement. This will not always be feasible, however, especially in national competitions for awards to conduct large-scale evaluations funded by such federal agencies as the National Institutes of Health (NIH), National Science Foundation (NSF), or U.S. Department of Transportation (DOT). The basic principle is that evaluators should take all feasible steps to consult with, or at least take into account the interests of, stakeholders prior to finalizing an evaluation agreement. Moreover, it is wise to provide for their inputs during a study.

Political Reasons for Evaluation Agreements

Another reason for negotiating advance evaluation agreements concerns the politics of evaluation. Evaluations can be intensely political, and political influences can impede an evaluation or even cause it to fail or be a party to unethical use of findings. Unchecked political forces can cause an evaluation to be unfair in its impact on program stakeholders. For example, in one of my (Daniel Stufflebeam's) evaluations, after completing the study, I learned that the client (the head of an organization) originally had no interest in obtaining information on the subject program's merit and worth. Instead, the client had contracted for the evaluation and used the evaluation report as a pretext for discharging the program's director, with whom he had a personality conflict. Although the report was decidedly positive, the client referenced the few indications of slight program inadequacy as the basis for firing the director. In another of my evaluations, from the start the client (a chief executive officer) had no intention of divulging to stakeholders any but favorable evaluation outcomes. Had I previously uncovered these clients' illicit reasons for requesting an evaluation, I would not have contracted to do the studies. These examples illustrate that an evaluator should, as feasible, examine carefully the politics surrounding a request for an evaluation before signing on to do the job (also see Patton [2008] on situational analysis).

Evaluators often need to make a wide search for the political forces that could have a negative influence on an evaluation. They should then institute relevant safeguards or, if appropriate and possible, decline the assignment. When feasible, evaluators should interview persons who might be harmed by the evaluation and give them every opportunity to express any concerns they have about the projected evaluation. Stakeholders who are vulnerable and in a position to be hurt by an evaluation often can alert a prospective evaluator to possibilities for illicit political influence.

Political threats can emanate from interest groups that want to bias, censor, or edit an evaluator's findings or even prevent the final evaluation report's release to rightful audiences.

Unfortunately, a client might have sought an evaluation as a cat's-paw for attacking an adversary or otherwise taking unfair advantage of stakeholders. Evaluators must be exceptionally careful not to become a tool of one side in a political dispute. Therefore, before finalizing an evaluation agreement and accepting an evaluation assignment, good practice is to search out and get input from a representative range of stakeholders. This sort of investigation can prove invaluable in deciding what to include in a proposed evaluation agreement and subsequently in deciding whether to proceed with an opportunity. Also, evaluators should take all feasible steps to adopt effective measures to ensure that political pressures will not interfere with or corrupt an evaluation during its execution. If sufficient safeguards cannot be instituted, an evaluator is wise to decline an opportunity, if that is possible. One safeguard we have often employed and found useful in militating against political interference in an evaluation is to engage and have regular exchanges with a review panel that is broadly representative of stakeholders.

We understand that evaluators who conduct studies in their own organization often cannot opt out of an assignment. They can, however, engage in relevant background work before formulating a recommended evaluation agreement. Moreover, by keying an agreement to professional principles and standards for evaluations (for example, American Evaluation Association, 2004; Joint Committee, 1994, 2011), they are in a strong position to argue their case for a sound evaluation agreement. As with external evaluations, reaching clear and professionally defensible understandings and written agreements with a client—about such matters as access to data, editing and release of reports, and use of findings—before starting a study is an important way to head off political threats to an evaluation. Also, the internal evaluator faced with a problematic assignment can provide the client with a printed presentation of the caveats under which the evaluation has to be conducted. In addition, he or she could ask the client to arrange for an independent metaevaluation based on appropriate professional standards.

Not all political influences are undesirable in evaluation work, however. By interacting with a representative group of stakeholders, an evaluator can build interest in and support for an evaluation. Stakeholders' involvement in an evaluation also can motivate them to consider and use evaluation findings. As noted earlier, it can be useful to appoint a stakeholder review panel and engage it throughout the evaluation to review and react to data collection plans and draft evaluation instruments and reports. Such a group can provide the evaluator with feedback of value in strengthening evaluation plans and materials. Further, in the process of conducting such reviews, panel members may become increasingly interested in and knowledgeable about evaluation procedures, and therefore become more enthusiastic about using an evaluation's findings. Also, panelists can become important opinion leaders, encouraging other stakeholders to take stock of and use an evaluation's findings and assisting them in doing so. Finally, members of such a panel often can facilitate the evaluator's collection of needed information.

Practical and Technical Reasons for Evaluation Agreements

Among the many practical and technical reasons for negotiating advance evaluation agreements are to establish clarity on deadlines; to establish protocols for entering program facilities and

collecting information from files and contacting human subjects; to allow for cooperation and support from personnel in the client organization; and to assign responsibility and authority in regard to disseminating findings. An evaluation design will have treated many of these items in detail. In contracting, it is important to make all such design items a matter of contractual agreement so that an evaluator can efficiently and effectively carry out the work with the approval and support of the client and other stakeholders. The evaluator should also include in the contract any important items that are not encompassed in the evaluation design and budget. Often it is prudent to append the evaluation's technical design to the body of the contract and stipulate that the technical design is part and parcel of the formal agreement.

Addressing Organizational Contracting Requirements

Of necessity, evaluators need to be familiar and involved with the grant-making and contracting practices of their organization and those of particular sponsors. Federal agencies (for example, NIH, NSF, and DOT); charitable foundations; and other sponsoring organizations often require that they and the evaluator's organization enter into a formal contract prior to launching an evaluation. In such cases, it is typical for attorneys in both the sponsoring organization and the contracting organization to become involved. We suggest that evaluators become acquainted with the contracting practices of their own organization and with those of potential sponsors. In our experience, attorneys can be useful in protecting their organization's interests and those of stakeholders, especially related to equal opportunity employment of evaluation team members, the rights of human subjects, deliverables, dissemination, liability protections, and payments. However, evaluators should not leave all the contracting to lawyers, who typically will not be sensitive to the full range of relevant methodological issues.

Negotiating Evaluation Agreements

The process of negotiating an evaluation agreement affords an evaluator and client an opportunity to base an evaluation effort on a constructive working relationship. It is in the interest of both client and evaluator to start their relationship in an atmosphere of mutual respect, confidence, and effective communication. Such a positive atmosphere is conducive to successfully addressing the many sensitive negotiation issues. In addition, the process of developing an evaluation agreement affords the client and the evaluator a valuable opportunity to review the evaluation design and budget; inform and obtain feedback from stakeholders concerning a semifinal draft of the contract; and ultimately clarify and agree on their individual and joint expectations, responsibilities, and rights in the evaluation.

As mentioned earlier, in addition to negotiating with the client, the evaluator should, as feasible, consult others who will be involved in, affected by, or interested in the evaluation but who are not parties to the written agreement. It would be a mistake to expect participation in the evaluation by persons who have not previously agreed to cooperate. When possible, such agreements for cooperation, for example, to assist data collection efforts, should be

obtained in advance, because they can be much harder to arrange later. When securing cooperation in advance is not feasible, the evaluator should explicitly include in the evaluation design and contract provisions for consulting stakeholders and obtaining their inputs and agreement to participate during a study. Among stakeholders to be contacted are those whose work will be assessed, those who will contribute information, administrators and staff in buildings where the evaluation will take place, members of the media, leaders of pertinent community organizations, and interested community members. We reiterate that the evaluator should, as feasible, seek out those who might be harmed by an evaluation so they can air any concerns they might have. If possible, all such consultations should occur before an evaluation agreement is signed. When this is not possible, the evaluator should ensure that these stakeholders have opportunities to present their views during the evaluation.

Prior to contracting, the evaluator will already have developed an evaluation design and budget. Whether in a formal request for proposal situation or a more informal sole-source case, the evaluator will present these items for the client's review and approval. Assuming that the evaluator's proposal is not rejected outright, the client typically will ask for some clarifications or changes.

Often this is where the negotiation process begins. The client and evaluator typically communicate about needs related to revising the evaluation design and budget, and the evaluator makes mutually acceptable changes. Subsequently either the client or the evaluator will prepare the first draft of the evaluation agreement for review and acceptance by both sides. If the evaluator prepares the first draft, he or she should have obtained stakeholder inputs if possible. If the sponsor has prepared the first draft, the evaluator should gather stakeholder input, possibly consult with an attorney, and desirably have an outside party review the draft for clarity and soundness. Ultimately this process should lead to a sound evaluation agreement—one that is keyed to professional standards, the evaluation design, and the budget; protects participants and other affected parties; reflects inputs from stakeholders; and ensures that the evaluation can be conducted efficiently and effectively.

Evaluation Contracting Checklist

Exhibit 21.1 is a checklist designed to help evaluators and clients identify key contractual issues and make and record their agreements for conducting an evaluation. Not all checklist items apply in every evaluation's set of agreements, as included in the evaluation contract or memorandum of agreement. It is prudent, however, to consider all of them when starting a negotiation and when reviewing a draft set of agreements. Then the parties to the contract or memorandum of agreement can select those items that should be incorporated into the instrument of agreement or revise the draft as appropriate. The evaluator can code each item as important and incorporated (indicated by a check mark) or as not applicable (NA); or he or she can leave it blank, indicating no agreement, although the item may be viewed as important. Mainly the checklist is a tool for evaluators to use in detailing and negotiating evaluation agreements; they can also sign, date, and retain the completed checklist as a convenient

summary of what they intended the agreement to cover. We will not elaborate or comment on the checklist items because we see them as self-explanatory.

Exhibit 21.1 EVALUATION CONTRACTING CHECKLIST

Basic Considerations

_____ Object of the evaluation (for example, a named program)
_____ Purpose of the evaluation
_____ Client
_____ Other right-to-know audiences
_____ Authorized evaluator(s)
_____ Guiding values and criteria
_____ Standards for judging the evaluation
_____ Contractual questions

Information

_____ Required information
_____ Data collection procedures
_____ Data collection tools
_____ Information sources
_____ Respondent selection criteria and process
_____ Provisions to obtain needed permissions to collect data
_____ Follow-up procedures to ensure adequate information
_____ Provisions for ensuring the quality of obtained information
_____ Provisions for storing and maintaining the security of collected information

Analysis

_____ Procedures for analyzing quantitative information
_____ Procedures for analyzing qualitative information

Synthesis

_____ Participants in the process of reaching conclusions
_____ Procedures and guidelines for synthesizing findings and reaching conclusions
_____ Decisions on whether evaluation reports should include recommendations

Reports

_____ Deliverables and due dates

_____ Formats for interim reports, including content, length, audiences, and methods of delivery

_____ Final report format, content, length, audiences, and methods of delivery

_____ Restrictions and permissions related to publishing information from or based on the evaluation

Reporting Safeguards

_____ Anonymity, confidentiality

_____ Prerelease review of reports

_____ Conditions for participating in prerelease reviews

_____ Rebuttal by evaluees

_____ Editorial authority

_____ Authorized recipients of reports

_____ Final authority to release reports

Communication Protocol

_____ Contact persons

_____ Rules for contacting program personnel

_____ Communication channels and assistance

Evaluation Management

_____ Timeline for evaluation work by the client and evaluators

_____ Assignment of evaluation responsibilities

Client Authority and Responsibilities

_____ Access to information

_____ Services (for example, clerical, office equipment, and telephone)

_____ Personnel

_____ Information

_____ Facilities

_____ Equipment

_____ Materials

_____ Transportation assistance

_____ Work space

Evaluation Budget

_____ Fixed-price, cost-reimbursable, or cost-plus agreement

_____ Payment amounts and dates

_____ Conditions for payment, including delivery of required reports

_____ Budget limits or restrictions

_____ Agreed-on indirect cost and overhead rates

_____ Contacts for budgetary matters

Review and Control of the Evaluation

_____ Contract amendment and cancellation provisions

_____ Provisions for periodic review, modification, and renegotiation of the design as needed

_____ Provision for evaluating the evaluation against professional standards of sound evaluations

Preparer: _____ Date: _____

Summary

This chapter has addressed the pivotal issue of evaluation contracting. An evaluation agreement may be a formal, legally enforceable contract or a less formal memorandum of agreement. In either case, the agreement should be grounded in professional standards for sound evaluations; keyed to the evaluation design and budget; eminently fair to all parties to the evaluation; oriented to ensuring the evaluation's feasibility; open to later amendment; preferably, written; and negotiated and signed in advance of starting the evaluation. Evaluators should negotiate sound advance agreements for their studies because these help prevent failure or abuses of an evaluation and also aid in securing stakeholder interest and cooperation. Sound contracting for evaluations is a process that begins with an evaluation design and budget and culminates in an agreement signed by the client and the evaluator. During the negotiation process, the evaluator is advised to obtain and take into account as much input as possible from the full range of stakeholders. We concluded the chapter by presenting a checklist of the full range of issues to consider when either developing or assessing an evaluation agreement.

Evaluations occur in the real world, where everything that can go wrong is likely to do so if left unchecked. Sound evaluation contracting provides a systematic approach to making an evaluation as fail-safe as possible.

REVIEW QUESTIONS

1. Provide two examples of how an evaluation client might use an advance evaluation contract to avoid an evaluator's later, possible misunderstandings about the evaluation's requirements. Then list two examples of how an evaluator could reference a well-constructed evaluation contract to address a client's attempts to discredit the evaluation or to persuade the client not to act improperly in regard to dissemination or use of the findings.
2. What are the definitions of memorandums of agreement and evaluation contracts, and what distinguishes them?
3. There are inherent dangers in a situation where either an evaluator or a client acts unilaterally in regard to aspects that should have been but were not included in an evaluation agreement. What are at least three such possible omissions? What could be the negative consequences of each omission?
4. At what points in an evaluation process should evaluation agreements be negotiated, and for what reasons?
5. Which of the following are legitimate signatories to an evaluation contract: a program beneficiary (for example, a student); a parent of a program beneficiary; the client; the head of the evaluation's stakeholder review panel; the evaluator; a newspaper reporter; the subject evaluand's accountant; or stakeholders who might be harmed as a consequence of the evaluation? Justify your reason for selecting each signatory. Then define each selected signatory's role in completing the contract. Finally, if you excluded any of the listed possible parties to the evaluation from being signatories to the evaluation contract, explain why they were excluded.
6. Identify and explain the importance of at least four issues—pertaining to a prospective client organization's request for an evaluation—that the evaluator should identify, examine, and resolve before finalizing an evaluation contract.
7. What are at least four types of possible political threats to an evaluation that evaluators should consider, and what do you see as potentially effective contractual precautions for defusing each identified threat?
8. What are at least three ways evaluators can make constructive use of political forces to enhance an evaluation's effectiveness?
9. What checkpoints might you include in a checklist for deciding whether or not to pursue a given evaluation?
10. Identify as many practical reasons for negotiating advance evaluation agreements as you can.

Group Exercises

Exercise 1

This chapter has emphasized the importance of establishing a constructive working relationship between evaluator and client. Discuss the following topics, from the perspective of the evaluator:

1. What factors make this relationship so vital?
2. How would you go about building such a relationship?
3. What is the role of a formal contract in making such a relationship sound and functional?
4. What are the hazards of developing too close of a relationship with a client?
5. Who should be involved before an evaluation agreement is concluded?
6. What are effective ways of involving these parties?
7. What are some of the factors that could preclude finalization of an evaluation contract?

Exercise 2

We have maintained that evaluation design, budgeting, and contracting are closely linked. Discuss this contention in general terms, and then select an actual evaluation known to a group member and find whether these three procedures were interrelated.

Exercise 3

Use the Evaluation Contracting Checklist in Exhibit 21.1 to role-play the development of an evaluation contract. Assign one member of the group to serve as convener and coordinator and to make assignments to the other group members, as follows:

1. Assign a second member to act as a potential evaluation client. Ask that client to find and bring to the group's next meeting a report of a completed evaluation and to be prepared to use the report as a basis for requesting a similar evaluation and negotiating a contract for its conduct.
2. Assign a third member to serve as the recorder of contractual agreements to be reached with the client at the group's next meeting.
3. Engage the additional members to serve as the evaluation team that is negotiating an evaluation contract with the client.

At the next meeting, begin by having the client describe her or his request for a program evaluation. Subsequently, the evaluation team should use the Evaluation Contracting Checklist to interview the client and reach tentative contractual agreements for the requested evaluation. Throughout this process, the meeting's recorder should document the contractual evaluation agreements tentatively being reached. At the end of the deliberation, the recorder should summarize the draft evaluation agreements. The client and evaluation team should then react to these and identify any needed changes or additions. Finally, the whole group should discuss

what was learned through the exercise, especially in regard to the utility of the Evaluation Contracting Checklist.

Suggested Supplemental Readings

American Evaluation Association Task Force on Guiding Principles for Evaluators. (2013). Guiding principles for evaluators. *American Journal of Evaluation, 34,* 145–146.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Joint Committee on Standards for Educational Evaluation. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

CHAPTER 22

COLLECTING EVALUATIVE INFORMATION

This chapter presents a detailed discussion of the concepts and techniques involved in collecting high-quality, reliable, and valid evaluative information.

After negotiating an evaluation agreement, the evaluator turns to the central task of collecting the needed information. Responses to a study's questions can be only as good and defensible as the supporting information. That information should address the full scope of questions developed through evaluator-client collaboration and possess sufficient depth. It should be reliable, appropriate, and credible. It also should be combinable into a coherent whole for reaching valid conclusions about an evaluand's merit and worth. The contract should reflect these intended outcomes. Collecting the needed information requires responsiveness to audience interests, technical competence, legal and ethical actions, human relations skills, meticulous management of information, and good measures of creativity and resourcefulness.

In this chapter we present practical advice on the information collection task. We begin by referencing the professional standards that are most relevant to collecting information. We then present a framework within which to plan information collection activities; it is intended to ensure sufficient scope and quality of obtained information and to encompass an evaluand's background, structure, operations, costs, and accomplishments. We end by summarizing selected information collection techniques that have proved especially useful in our evaluations.

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **Which evaluation standards pertain to the core issues in collecting evaluative information**
- **Checklists of practical steps for conducting each part of the information collection process**
- **Alternative sampling approaches**
- **Different types of reliability**
- **Different types of validity**
- **Requirements for managing information**
- **A framework for planning data collection procedures**
- **Alternative techniques for collecting evaluative information**

Key Standards for Information Collection

The evaluation discipline includes clear standards; a vast literature base; and strong procedures for developing, validating, and employing tools for collecting information.

The Joint Committee on Standards for Educational Evaluation (2011) defined a range of standards to be met by an evaluation's information collection devices and activities: Relevant Information, Human Rights and Respect, Explicit Program and Context Descriptions, Reliable Information, Valid Information, and Information Management. Moreover, the literature pertinent to meeting these standards is extensive. It includes guidelines and procedures for developing instrument blueprints, constructing response items, drafting and pilot-testing instruments, performing item analysis, finalizing instruments, developing norms, performing reliability and validity studies, selecting appropriate samples of respondents, controlling information collection conditions, verifying obtained data, keeping collected information secure, attending to a range of ethical considerations, and many more. To the extent that evaluators can find or develop instruments—both quantitative and qualitative—that pass muster within the canons of valid assessment, they are in a strong position to collect and defend the information needed to evaluate a program or other evaluand. By using a sufficiently broad range of pertinent and defensible information, evaluators can answer evaluative questions confidently, reach conclusions about an evaluand's quality and accomplishments, and persuasively advocate use of an evaluation's findings.

Therefore, evaluators need a firm grasp of standards for collecting sound evaluative information and should attend as carefully and systematically as they can to meeting these standards. In this section we review what we consider to be the most important Joint Committee (2011) standards related to collecting evaluative information. For each standard, we present its summary statement, explain it, add our comments, and note its relevance to the collection of evaluative information. We advise readers to consult the full text of each standard in the standards document itself.

Before proceeding, a caveat is needed. The practical constraints of many evaluation assignments preclude meeting—or make it extremely difficult to fully meet—the standards associated with sound information collection. Evaluators often cannot access preexisting instruments that are valid for collecting information needed to draw inferences in a particular evaluation. In addition, they rarely have sufficient time and resources to carry out a full sequence of developing, pilot-testing, reformulating, norming, and validating new instruments, a process that could require years of painstaking and expensive work. Typically, competent evaluators construct new instruments as systematically as they can; use other available instruments that may have only marginal validity for the particular study; gather existing relevant, reliable information; and repeatedly replicate the information collection process until findings are stable and defendable. In proceeding in this manner, evaluators need to take reasonable steps to meet the standards related to information collection. They should not, moreover, expect users of the evaluation to be comfortable or completely satisfied with information collection procedures and tools that fall short of fully meeting the principles of sound inquiry. Especially, evaluators should report to users any deficiencies in their collected information. Among the

steps an evaluator should take to overcome weaknesses in information collection instruments is to build in cross-checks on findings. The main way to do this is to employ multiple sources of information and multiple methods, look for consensus in findings, and report both discrepancies and agreements in the findings. In sum, evaluators should take the standards seriously and do all they can to meet them, and they must forthrightly report limitations in the presented information. From this perspective, evaluators' modest aims are to at least increase the rationality and defensibility of their conclusions, judgments, and decisions (Bamberger, Rugh, & Mabry, 2012).

Relevant Information

The standard states, "Evaluation information should serve the identified and emergent needs of intended users" (Joint Committee, 2011, p. 45). The explanation of this standard stresses two main requirements: scope and selectivity. To meet the spirit of the standard, evaluators often should provide both more and less information than the client requests.

Evaluators should collect information that has sufficient scope to address an audience's most important information needs and support a judgment of merit and worth. Typically evaluators should obtain information on all the important variables (for example, participant needs, program goals and assumptions, program design and implementation, program costs and outcomes, and positive and negative side effects). They should collect information on all the essential questions, whether or not the client and stakeholders specifically request the information. This is in keeping with the dictum that an evaluation is not just an information service, but essentially an assessment of merit and worth. As the Joint Committee (1994) stated, "Evaluators should determine what the client considers significant but should also suggest significant areas the client may have overlooked, including areas identified by other stakeholders" (p. 37). Later in the chapter, we present a framework to help evaluators and clients ensure that they at least consider a comprehensive set of possible assessment variables.

The second important aspect of this standard is that evaluators necessarily have to be selective in deciding what information to collect. Typically it is not feasible to satisfy all the information interests of all stakeholders. However, not all of the contemplated information will be equally important to stakeholders or essential for reaching evaluative conclusions. Initially, an evaluator should identify the potential body of relevant information, including the information that the client and stakeholders desire and also what is needed to render a judgment of merit and worth. Subsequently, the evaluator should work with the client and stakeholders to separate the most important items of information from those that are only desirable or of minor importance. Then, in consideration of available funds and time, the evaluator should select the information to be collected judiciously.

Exhibit 22.1 is a checklist of actions of use in meeting the central requirements of the Relevant Information standard.

Exhibit 22.1 RELEVANT INFORMATION CHECKLIST

Checkpoints

_____ Interview stakeholders to determine their different perspectives; information needs; and views concerning what constitutes credible, acceptable information.

_____ Plan to obtain sufficient information to address the client group's most important information needs.

_____ Assess and adapt the information collection plan to ensure adequate scope for assessing the program's value (for example, its worth, merit, and/or significance).

_____ Ensure that the obtained information will address and keep within the boundaries of the evaluation's stated purposes and key questions.

_____ Allocate time and resources to collecting different parts of the needed information in consideration of their differential importance.

_____ Allow flexibility during the evaluation process for revising the information collection plan pursuant to the emergence of new, legitimate information needs.

Human Rights and Respect

The standard states, "Evaluations should be designed and conducted to protect human and legal rights and maintain the dignity of participants and other stakeholders" (Joint Committee, 2011, p. 125).

Most program evaluations involve gathering information from and pertaining to a wide range of persons associated with the subject program: program beneficiaries, staff, administrators, policymakers, community members, and others. Without due process and care, evaluators might unwittingly or otherwise violate such persons' rights. These violations can embarrass or harm the affected persons, evoke legal prosecution or professional sanctions, stir up dissension, and discredit the evaluation and render it ineffective. Evaluators should thus systematically identify and make provisions for adhering to all applicable rights of those who are parties to the evaluation. Among the human rights to uphold are those concerned directly with the persons' roles in the evaluation and a wide range of more pervasive rights. Some rights are based in law, whereas others derive from ethics and common courtesy. The Joint Committee (1994) noted,

> Legal provisions bearing on rights of persons include those dealing with consent for participation, privilege of withdrawal without prejudice and without withdrawal of treatment or services, privacy of certain opinions and information, confidentiality of information, and health and safety protections. (p. 93)

In addition, ethical, commonsense, and courtesy considerations require evaluators to honor evaluation participants' right not only to place limits on the extent and timing of their involvement in information collection but also to decline experiences they consider to be

detrimental or uncomfortable. Evaluators must also respect the cultural and societal values of all participants (Morris, 2003).

One important way to uphold the rights of human subjects is to have an appropriate human subjects institutional review board vet one's evaluation design. Another is to strictly follow the advice in the full text of the Human Rights and Respect standard (Joint Committee, 2011). Accordingly, evaluators should consider developing and sharing with human subjects the procedures that they and the client will follow to ensure that participants' rights are protected. Evaluators should inform prospective information providers how the information they provide will be used. They should secure written permission from duly authorized parties to access individual records (if needed), and they should make every effort to protect the identity of those who respond to evaluation instruments or otherwise supply information. They should, where applicable, obtain permission from respondents and the client to tape-record interviews. Although it is often desirable to provide confidentiality or anonymity in gathering information, the evaluator should not promise either one when it cannot be guaranteed. It is also desirable in some evaluations to make special provisions to enable language-minority participants to supply information, even if this means translating instruments into their first language. When minors are involved in program evaluation (as occurs, for instance, with school-based studies), parental permission must be sought and given, and often an appropriate adult should be present during an interview session.

Shown in Exhibit 22.2 is a checklist of actions of use in meeting the main requirements of the Human Rights and Respect standard.

Exhibit 22.2 HUMAN RIGHTS AND RESPECT CHECKLIST

Checkpoints

_____ Adhere to applicable federal, state, local, and tribal regulations and requirements, including those pertaining to institutional review boards, local or tribal constituencies, and ethics committees that authorize the conduct of research and evaluation studies.

_____ Take the initiative to learn about, understand, and respect stakeholders' cultural and social backgrounds, local mores, and institutional protocols.

_____ Make clear to the client and stakeholders the evaluation's provisions for adhering to ethical principles and codes of professional conduct, including the standards of the Joint Committee (2011).

_____ Institute and observe rules, protocols, and procedures to ensure that the evaluator, or all evaluation team members, will develop rapport with and consistently manifest respect for stakeholders and protect their rights.

_____ Make stakeholders aware of their right to participate, withdraw from the study, or challenge decisions that are being made at any time during the evaluation process.

_____ Monitor the interactions of evaluation team members and stakeholders and act as appropriate to ensure continuing, functional, and respectful communication and interpersonal interactions throughout the evaluation.

Explicit Program and Context Descriptions

The standard states, "Evaluations should document programs and their contexts with appropriate detail and scope for the evaluation purposes" (Joint Committee, 2011, p. 185).

When a final evaluation report is issued, readers need to know what was evaluated. It is insufficient to describe the program only as it was originally conceived, because its implementation may have been quite different. Also, it is not enough to characterize the program only in general terms, because many readers need details. Readers who are interested in replicating the program require sufficient particulars to contrast the program with critical competitors. If they decide to adopt the program, they need specific information to help decide how to organize their version of the program, finance it, launch it, and make it work. When a program fails, program funders and administrators need information on program expenditures, staffing, and operations to diagnose the reasons for failure. Moreover, researchers who want to understand a program's effects need detailed information about the program's actual operations so they can relate parts of the program to its outcomes.

An evaluator should collect sufficient information to help members of the audience understand both the program's original plan and its actual implementation. Clearly the evaluator should collect information on how the program was structured, governed, staffed, financed, and carried out; where it was conducted; what facilities were used; what orientation and training participants received; how much community involvement (if any) took place in the program; and how program funds were budgeted and spent. Relevant sources of extant information about the program might include generalized program descriptions, funding proposals, public relations reports, minutes of staff meetings, media presentations, newspaper accounts, expense reports, and progress and final reports. Additional information might need to be collected from participant observers and independent observers, from interviews with various program participants, from focus groups, and from direct observation by the evaluator. Photographic records can also prove enlightening. In collecting information, it is wise to obtain both holistic descriptions and descriptions of program components. Over time, it can be useful to record time-specific descriptions to document and contrast changes in the program and identify trends. It is especially important to search out discrepancies between intended and actual program operations. Program documentation is a major information collection task.

This standard also calls for studying and documenting the context in which the program exists, so that its likely influences on the program can be identified. We have seen repeatedly that a program's context can heavily influence how the program is designed and operated and what it achieves. Two or more programs with the same design often differ considerably in implementation and outcomes due to the influences of their respective backgrounds and environmental circumstances. To understand how a program acquired its particular characteristics and why it succeeded or failed, an evaluator needs to collect considerable contextual information. Important contextual variables include the program's

geographical location, the relevant political and social milieu, the economic health of the surrounding community, program-related needs and problems in the area, pertinent legislation, availability of special funds for work in the program area, highly influential persons, highly influential environmental events, the program's rationale and means of getting started, its organizational home, its timing, its potential contributions to the locale, actual participants and their program-related needs, competing programs in the area, and pertinent state and national influences.

Formative evaluations require contextual information to help those in charge of a program take account of local circumstances and identify and address participants' needs and problems. Summative evaluations require contextual information to help audiences understand why a program succeeded or failed and to prevent erroneous interpretations about how applicable the findings are to other contexts.

To consider how a program might work elsewhere, audiences need to know what highly influential contextual dynamics would have to be present in that new setting. For example, an audience might want to know whether a program's success had been aided heavily by a supportive community, a charismatic and unusually effective political leader, a sizable subsidy from a local foundation, or a history of social service agency cooperation. From another point of interest, an audience still might want to consider replicating an unsuccessful program if its failure clearly was due to environmental circumstances beyond the program's control. Examples of such negative influences could be local unrest, weak leadership and direction, a fiscal crisis in the program's organization, high staff turnover, or area devastation by a hurricane or tornado.

In describing a program's context, an evaluator should draw information from multiple sources. Such information might include minutes of board meetings, news accounts, area demographic statistics, area economic data, and pertinent legislation, for example. Also, evaluators are advised to maintain a log of unusual circumstances. Negative forces might include a destructive flood, a strike in the program's organization, departure of a major area corporation, embezzlement of program funds, civil unrest, unanticipated changed legislation, unexpected opposition from area interest groups, or a health epidemic. Unexpected positive influences could be a major corporation's move to the area, an unanticipated grant from a national philanthropic foundation, or a scientific discovery. Relevant contextual information can be obtained from a wide range of individuals and groups, the local chamber of commerce, members of fraternal organizations, local clergy, real estate agents, newspaper reporters, corporation officials, local charitable foundation officials, social service organizations, law enforcement officials, court officials, and area businesspeople. It is often highly desirable to compile a photographic or video record of key aspects of the program's setting.

Shown in Exhibit 22.3 is a checklist of actions of use in meeting the main requirements of the Explicit Program and Context Descriptions standard.

Exhibit 22.3 EXPLICIT PROGRAM AND CONTEXT DESCRIPTIONS CHECKLIST

Checkpoints

_____ Describe all important aspects of the program (for example, goals, design, intended and actual recipients, components and subcomponents, staff and resources, procedures, and activities) and how these evolved over time.

_____ Describe how people in the program's general area experienced and perceived the program's existence, importance, and quality.

_____ Identify any model or theory that program staff invoked to structure and carry out the program.

_____ Define, analyze, and characterize contextual influences that appeared to significantly influence the program and that might be of interest to potential adopters, including the context's technical, social, political, organizational, and economic features.

_____ Identify any other programs, projects, or factors in the context that may have affected the evaluated program's operations and accomplishments.

_____ As appropriate, report how the program's context may be similar to or different from other contexts in which the program is expected to be adopted or might reasonably be adopted.

Defensible Information Sources

Here we reference the Defensible Information Sources standard from the Joint Committee's 1994 edition of *The Program Evaluation Standards*. We do so because accessing appropriate, often multiple sources of information and employing appropriate sampling methods are critically important in the process of producing sound evaluation reports and because the Joint Committee's 2011 edition includes no clearly identifiable standard on defensible information sources.

The standard states, "The sources of information used in a program evaluation should be described in enough detail, so that the adequacy of the information can be assessed" (Joint Committee, 1994, p. 141).

Because evaluation is mainly a time-constrained enterprise that functions under real-world complexities, evaluators typically should employ a variety of techniques to collect information from multiple sources. Sources may include program staff and participants, administrators and policy board members, newspapers and public records, program proposals and reports, and program records. Evaluators employ a variety of techniques to tap these sources. Persons may be surveyed, tested, interviewed, engaged in focus groups or hearings, or asked to complete a

rating scale. Documents may be coded and subjected to content analysis. Program activities may be observed or photographed. Evaluators typically should employ both qualitative and quantitative methods. By using a variety of techniques to obtain information from multiple sources, evaluators provide cross-checks and perspective for addressing each major evaluation question and help ameliorate limitations of individual sources and methods.

Evaluators often cannot collect all of the potentially relevant information from each source of information. For example, they cannot observe, test, and interview every participant during every day of a program. No one would expect or tolerate such an extreme quest for information. In selecting information sources and methods, evaluators should avoid overloading respondents with unreasonable requests for information. Evaluators often collect information from only a sample of participants and only on a few days in the program. Moreover, they may have time to examine only a sample of records and other documents. Consequently, they need to be selective in collecting information.

Such selectivity introduces possibilities for both biased and missing information. If the information is not representative of what occurred in the program or of responses that might have been obtained from all members of a particular respondent group, the evaluator may draw erroneous conclusions and mislead an audience. Therefore, he or she needs to introduce appropriate safeguards to enhance representativeness and transparency in his or her findings. The evaluator should be forthright in reporting limitations of his or her information sources. The Joint Committee (1994) advised evaluators to "document, justify, and report their sources of information, the criteria and methods used to select them, the means used to obtain information from them, and any unique and biasing features of the obtained information" (p. 141). The committee noted further that "poor documentation and description of information sources can reduce an evaluation's credibility" (p. 142). An evaluation's technical appendix (or a separate technical report) should include documentation of information sources, the information collection process, and the instruments used to collect the information.

In most studies, an evaluator selects samples (including program documents) from pertinent populations using a wide range of sampling techniques, including both probability and nonprobability sampling methods. The objective of probability sampling is to estimate population parameters from information contained in a sample—that is, to make inferences about a population from information contained in a sample selected from that population. In most instances, such inferences are in the form of an estimate of a population parameter, such as a mean, total, or proportion, with a bound on the error of estimation (sometimes referred to as the margin of error). Each observation taken from a population contains a certain amount of information about the population parameter or parameters of interest. Therefore, the central feature of nearly all sampling designs is to determine the necessary sample size or quantity of information in a sample pertinent to a population parameter. The essential nomenclature related to sampling includes the terms *element, population, sample, sampling unit,* and *frame*. An element is an object on which a measurement is taken. A population is

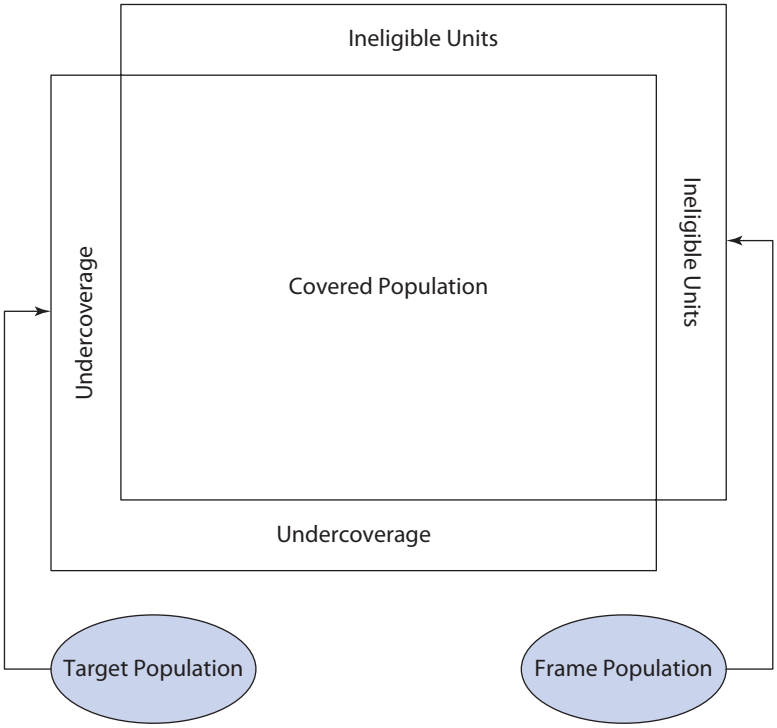

Figure 22.1 Coverage of a Target Population by a Sampling Frame

Source: Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley, 54.

a collection of elements about which an inference is made based on a sample. A sample is a collection of sampling units drawn from a frame or frames, often called the population of interest. Sampling units are nonoverlapping collections of elements from the population that cover the entire population. A frame is a list of sampling units (see Figure 22.1). In the figure, overcoverage occurs when units that are not part of the target population or sampling frame are included in a sample, and undercoverage occurs when units that should be part of the target population or sampling frame are excluded. Both overcoverage and undercoverage create sampling errors, which ultimately may result in incorrect inferences made from a sample to a target population.

When probability sampling methods (and nonprobability sampling methods, in some instances) are used, evaluators should expend substantial effort to reduce coverage, sampling, nonresponse, and measurement errors (Dillman, Smyth, & Christian, 2009). Coverage errors are those that arise when not all members of a population have a known, nonzero probability of selection and when those who are excluded are different from those who are sampled. Sampling errors reflect the extent to which a sample estimate differs from a population parameter because not every unit in the population is sampled. Nonresponse errors occur when units selected

who do not respond are somehow different from those who do. In many sampling situations, nonresponse is one of the most pervasive problems, and evaluators would be wise to apply Dillman et al.'s social exchange theory approach (2009) to reduce nonresponse. Measurement errors occur as a result of inaccurate or imprecise responses.

A stratified random sample is one obtained by separating the population elements into discrete, nonoverlapping groups (for example, males and females), called strata, and then selecting a simple random sample from each stratum. The principal reasons for using stratified random sampling rather than simple random sampling are as follows:

- Stratification may produce a smaller bound on the error of estimation than would be produced by a simple random sample of the same size. This is particularly true if measurements within strata are homogeneous.
- The cost per observation may be reduced by stratification of the population elements into convenient groupings.
- Estimates of population parameters may be desired for subgroups of the population. These subgroups should then be identifiable strata.

Cluster sampling is a less costly alternative to simple or stratified random sampling if the cost of obtaining a frame that lists all population elements is very high or if the cost of obtaining observations increases as the distance separating elements increases. Cluster sampling is an effective design for obtaining a specified amount of information under the following conditions:

- A good frame listing all population elements is not available or is very costly to obtain, but a frame listing clusters is easily obtained.
- The cost of obtaining observations increases as the distances separating the elements increases.

Clusters typically consist of herds, households, or other units of clustering. For example, a farm herd contains a cluster of livestock for estimating proportions of diseased animals. Elements within a cluster are often physically close together and hence tend to have similar characteristics, and the measurement on one element within a cluster may be correlated with the measurement on another. The quantity of information contained in a cluster sample is affected by the number of clusters and the relative cluster size.

Notably, most statistical theory is premised on an underlying infinite population. By contrast, sampling theory and practice are founded on the assumption of sampling from a finite population, as is often the case in evaluation scenarios in which a program serves a fixed, often relatively small population. In the general framework of finite population sampling, samples of size n are taken from a population of size N (that is, a population with N elements or members). In the finite population case, the variance estimate of a statistical estimator, such as a mean or total, must be adjusted using the finite population correction (fpc), due to the fact that not all

data from a finite population are observed. For simple random samples (without replacement), the fpc is expressed as

$$1 - \frac{n}{N} \quad \text{or} \quad 1 - f$$

where f is the sampling fraction or the sampling rate:

$$f = \frac{n}{N}$$

The fpc is, therefore, the fraction of a finite population that is not sampled. Because the fpc is literally a factor in the calculation of an estimate of variance for an estimated finite population parameter, the estimated variance is reduced to zero if $n = N$. In samples of very large populations, f is very small, and the fpc may be ignored. Although the fpc is applicable for estimation, it often is not necessary for many inferential uses, such as in statistical significance testing (for example, comparison between sampled subgroups). For detailed information on probability sampling methods, including how to select an appropriate sample size, see Cochran (1970); Henry (1990); Kish (1965); Koleci, Coryn, Hobson, and Keci (2011); Lohr (2010); and Scheaffer, Mendenhall, Ott, and Gerow (2012).

Although probability samples generally are advocated in the evaluation literature, in practice evaluators employ a variety of other sampling approaches with beneficial results. They may employ such nonprobability sampling methods as purposive sampling to obtain information from key informants, such as a policy board's chair, a program's director, a program's task leader, or a program's internal evaluator. In many evaluations, it is essential to obtain information from such stakeholders, and probability sampling would not be applicable. Another class of especially useful methods comprises the various types of chain-referral sampling methods, such as snowball sampling (that is, interviewing one or a few individuals at the outset and asking each to identify others who should be interviewed) and respondent-driven sampling, which work well for certain hard-to-reach and hidden populations, such as the homeless, injection drug users, and HIV-positive individuals (Coryn, Gugiu, Davidson, & Schröter, 2008). An advantage of these approaches is that they can guide evaluators to key informants who otherwise might not have been sampled. Patton (2002) described fifteen sampling designs (including the aforementioned chain-referral sampling method) for use with qualitative evaluation approaches. Of particular relevance to many evaluations are the following approaches he recommended:

- *Extreme or deviant case sampling.* Units are selected because they are in some way unusual or deviant (for example, success cases, failure cases).
- *Intensity sampling.* Selected units manifest the phenomena or phenomenon of interest intensely (but not in as extreme a manner as those considered deviant cases).
- *Homogeneous sampling.* Units are selected from subgroups so as to describe each subgroup in depth.
- *Typical case sampling.* Units are selected that manifest typical characteristics of interest.

- *Critical case sampling*. Selected units are considered particularly important.
- *Criterion sampling*. Units are selected on the basis of predetermined criteria (for example, on the basis of exposure to a critical incident of interest).

Regardless of an evaluation's sampling plan, interested stakeholders who were not sampled may desire to provide their input. There are good reasons for the evaluator to accept their information and incorporate it in the evaluation. For one thing, doing so lends credibility to the study and can influence the volunteers to take study findings seriously. Also, the evaluator may learn unique and valuable lessons from the volunteers. The main caveat here is that the evaluator should keep the volunteered information separate from the other obtained information, analyze the different sets separately, and inform readers of the limitations of such volunteer samples.

Not all evaluations have to involve sampling of respondents. In some studies, the evaluator obtains information from all of a program's participants, each staff member, and each member of a policy board (that is, taking a census rather than selecting a sample). When data are gathered from all members of a population, the results can be reported directly. In such cases, there is no need to make inferences about the population based on a sample because the evaluator has drawn information from the total population. Doing this simplifies considerably an evaluator's tasks of analyzing and reporting findings. In planning data collection activities, it often is appropriate to consider the feasibility and desirability of taking measures from all members of a population of interest. If this can and should be done, the evaluator proceeds to take population measures and sets aside concerns about sampling. Otherwise, the evaluator should select and apply appropriate sampling techniques.

Whether they employ sampling or a population census approach, evaluators should report the information selection experience forthrightly, including its nature and its strengths and weaknesses. They should describe the sources of information, document the techniques and processes by which information was collected from each source, and document changes and difficulties that occurred along the way. When information was gathered through a cascading or iterative process (for example, successively drawing from a sequence of unfolding events or interactions), an evaluator should report and justify the rules he or she followed to decide when to cease collecting information (for example, because of redundancy, new information that was only of marginal importance, or lack of additional time or resources). In regard to information that was collected according to a prespecified plan, an evaluator should report the original plan, any deviations from the plan, and the import of the deviations for interpreting findings.

Evaluators typically obtain both qualitative and quantitative information. They should not automatically value one type of information more than the other, but should report the strengths and weaknesses of each and also their complementary nature. Again, we stress that evaluators should document and report both strengths and deficiencies in their information sources and, as appropriate, caution the audience not to place undue confidence in the obtained information.

In summary, Exhibit 22.4 shows a checklist of key requirements for meeting the Defensible Information Sources standard.

Exhibit 22.4 DEFENSIBLE INFORMATION SOURCES CHECKLIST

Checkpoints

_____ Obtain information from a variety of sources.

_____ Use pertinent, previously collected information once validated.

_____ As appropriate, employ a variety of qualitative and quantitative data collection methods.

_____ Document and report sampling designs and procedures.

_____ Document and report any biasing features in the obtained information.

_____ Include sampling plans and data collection instruments in the evaluation report's technical appendix (or in a separate technical report).

Reliable Information

The standard states, "Evaluation procedures should yield sufficiently dependable and consistent information for the intended uses" (Joint Committee, 2011, p. 179).

Reliability is a necessary but not sufficient condition for validity (Crocker & Algina, 2008; Nunnally & Bernstein, 1994). An evaluative conclusion cannot be defended as valid if it is based on unreliable information. Information is unreliable to the extent that it contains unexplained contradictions and inconsistencies or if different answers would be obtained under subsequent but similar information collection conditions, absent a known intervention. In the parlance of classical test theory (CTT), reliability is the consistency or reproducibility of scores. In the context of CTT, there are two types of measurement error:

- *Random measurement error*. This type of error consistently affects an individual's score because of purely chance happenings. Random measurement error may affect an individual's score in both positive and negative directions, thus cancelling out in the long run. Examples include blind guessing, administration errors, scoring errors, and distractions during testing.
- *Systematic measurement error*. This type of error consistently affects an individual's score because of some particular characteristic of the person or test that has nothing to do with the construct being measured. Systematic measurement errors tend to accrue. Examples include biased raters, scoring key errors, and examinee test anxiety.

In CTT, the objective of any measurement procedure is to identify a person's (or other object's) true score, which is expressed as

$$X = T + E$$

where X represents an observed score, T represents the true score, and E represents measurement error (Guilford, 1936; Lord & Novick, 1968). Relatedly, measurement error is the discrepancy between an examinee's observed score and his or her true score, or more formally (where the subscript j represents a random examinee),

$$E_j = X_j - T_j$$

Reliability coefficients derived from this logic can range from 0.00 to +1.00, where a reliability coefficient of 0.00 indicates that all measurement variation is attributed to error, whereas a reliability coefficient of +1.00 indicates no measurement error. The closer a reliability coefficient is to +1.00, the more confident an evaluator can be that a measurement is an accurate representation of a person's true score. So, for example, if a reliability coefficient were 0.85, then 85 percent of the variance is due to variability in true scores, whereas 15 percent is error variance. Under CTT, reliability estimates, usually in the form of correlation coefficients, can be made under the following conditions:

- *Test-retest (stability).* The correlation between the same examinees is tested on different occasions. Error reflects random fluctuations in performance over time.
- *Alternative forms (equivalence).* The correlation between the same examinees is tested with different tests on the same occasion. Error reflects random fluctuations in content (item) sampling.
- *Internal consistency (Cronbach's α, split-half).* There is agreement between performance on individual items and overall performance on the total test. Error reflects content (item) sampling and heterogeneity of the behavioral domain sampled.

Evaluators can be concerned about one or more of these types of reliability, depending on the nature of an evaluation. Reliability can be influenced by the variability of a sample, where homogeneous samples tend to lower reliability and heterogeneous samples tend to increase reliability; test length; and the time limit of a test, among others. CTT has many conceptual and statistical problems, however, such as that (1) most reliability coefficients are based on correlation coefficients, which do not measure reliability per se, but rather the covariance or rank order among a set of measurements; (2) reliability coefficients are a function of sample characteristics, and the same measuring device will produce different reliability coefficients in different samples; (3) most correlation coefficients are bivariate statistics, so only two sets of scores can be examined at the same time; and (4) in CTT, the error term, *E,* cannot be partitioned into random measurement error and systematic measurement error components.

Later developments in psychometric and measurement theory, such as generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) and item response theory (IRT), however, have resolved many of the conceptual and statistical problems associated with CTT. Detailed presentations of generalizability theory and IRT exceed the scope of this book, and interested readers are referred to Brennan (2001), de Ayala (2009), and Embretson and Reise (2000).

Many evaluations involve the use of raters or observers to gather information about a program, such as when raters or observers use checklists to verify—through direct observation—that a program is being implemented with fidelity, when they employ rubrics to score or grade written work samples of program participants, or when they apply coding dictionaries to code qualitative information obtained through interviews or focus groups. When multiple raters, observers, or coders are used, one or more interrater reliability coefficients, such as the simple coefficient of agreement, p_o; Cohen's κ to account for chance agreements; or various models of intraclass correlation coefficients can be used to assess the extent to which raters, observers, or coders provide consistent estimates about what they observe, rate, code, or judge (see Davey, Gugiu, and Coryn [2010] for a comparison of interrater reliability coefficients and formulas for calculating them).

Evaluators should determine which forms of reliability are most applicable to their study and make appropriate assessments. For all forms of reliability, they should strive to reduce or document the amount of error variance and its impact on an evaluation's information and conclusions.

Shown in Exhibit 22.5 is a checklist of actions evaluators can take to ensure that their evaluative conclusions meet the Reliable Information standard.

Exhibit 22.5 RELIABLE INFORMATION CHECKLIST

Checkpoints

_____ Determine, justify, and report the needed types of reliability—test-retest, interrater reliability, or internal consistency, for example—and acceptable levels of reliability.

_____ In the process of examining, strengthening, and reporting reliability, account for situations in which assessments are or may be differentially reliable due to varying characteristics of persons and groups in the evaluation's context.

_____ Ensure that the evaluator, or evaluation team, possesses or has access to the expertise needed to investigate the applicable types of reliability.

_____ Describe the procedures used to achieve consistency (for example, between raters or observers).

_____ Provide appropriate reliability estimates for key information summaries, including descriptions of the program, program components, context, and outcomes.

_____ Examine and discuss the consistency of scoring, categorization, and coding between different sets of information (for example, in assessments by different observers).

_____ When choosing from extant instruments, select ones that previously yielded information with acceptable reliability for answering questions like those in the projected evaluation.

_____ Clearly determine the unit of analysis for each information collection device, and assess reliability at that level of discourse.

_____ Carefully develop and follow a blueprint in constructing each information collection device, to include its rationale, target questions, sources of information, means of administration, and appropriate form of reliability assessment.

_____ Draft, pilot-test, and refine all new instruments.

_____ Engage stakeholders to review draft instruments and draft sets of findings, and carefully consider and address their assessments.

_____ Depending on the size of the evaluation, engage multiple data collectors, and examine their findings for consistency.

_____ Systematically train data collectors and those who will code, score, and analyze obtained information.

_____ Document procedures to strengthen and assess reliability and results in the evaluation report's methods section or technical appendix (or in a separate technical report).

Valid Information

The standard states, "Evaluation information should serve the intended purposes and support valid interpretations" (Joint Committee, 2011, p. 171).

Validity, in general, is considered the approximate truthfulness or correctness of an inference or conclusion. More specifically, validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationale support the adequacy and appropriateness of inferences and actions based on some form of assessment. Therefore, validity or truth consists of two equally important premises, one inferential and the other consequential (Messick, 1989). Even so, validity is not an all-or-none proposition. It is one of degree.

Four classes of validity are of central concern in most evaluations, though there are many others (see Figure 22.2). Two of these, internal and external validity, are largely, but not specifically, functions of a study's design. Internal validity concerns the truthfulness or correctness of inferences about whether the relationship between two or more variables is causal. External validity concerns the extent to which a causal relationship holds over variations in persons, settings, treatments, and outcomes, and is largely a function of sampling. Construct validity, in contrast, is primarily a function of measurement. Construct validity concerns the degree to which an inference about an observed person, setting, treatment, or set of outcomes is warranted given the soundness of the employed theoretical framework, defined population of interest, sampling design, instruments and procedures used for measurement, obtained information, and analysis and synthesis procedures. Finally, conclusion validity, sometimes referred to as statistical conclusion validity, is predominately a function of analysis. Conclusion validity concerns the correctness of inferences regarding the covariation between two or more variables. Although conceptually independent, these types of validity are not mutually exclusive. Taken mainly from the quantitative tradition, they are nonetheless found in both

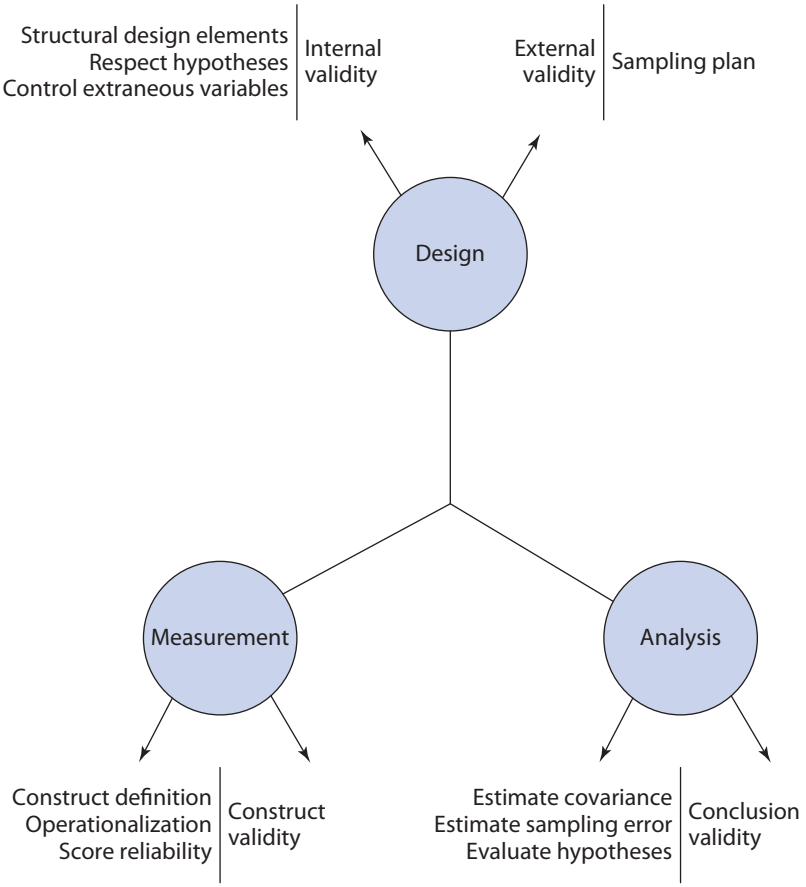

Figure 22.2 Types of Validity Addressed by Design, Measurement, and Analysis

Source: Kline, R. B. (2008). *Becoming a behavioral science researcher: A guide to producing research that matters*. New York, NY: Guilford Press, 40.

the quantitative and qualitative traditions, though the terminology used to represent these concepts differs somewhat. Kline (2008) has provided an accessible introduction to these concepts and how each can be addressed in a study.

Among the possible information-gathering products and associated processes are results of interviews, observations, document reviews, hearings, forums, focus groups, testimony, surveys, and administration of performance tests or objective tests. The processes should be chosen and employed to produce information that is relevant to a study's questions, reliable, and sufficient in scope and depth to answer all of those questions.

Validation of instruments and procedures is required to ensure the soundness of the obtained information for answering a study's questions. According to the Joint Committee (1994), "Validation is the process of compiling evidence that supports the interpretations and uses of the data and information collected using one or more . . . instruments and procedures" (p. 145).

Following are the tasks in a sound process to validate a given information collection instrument or procedure:

- Provide a detailed description of the program attribute about which information is required; examples are the program's context, design, implementation, and outcomes.
- Determine the type of information—for example, a description or judgment—the particular information collection instrument or procedure is intended to acquire.
- Determine the type of information the tool or procedure provides.
- Describe in detail how the tool or procedure was applied and how well its application was monitored and controlled.
- Describe in detail and assess the credibility of the persons who collected or supplied the study's information.
- Determine the appropriate unit of analysis.
- Analyze the reliability of the obtained information—that is, its consistency and/or reproducibility, and judge whether reliability is sufficient for the intended use.
- Describe in detail and assess the procedures used to score, code, analyze, and interpret the obtained information.
- Compile qualitative and quantitative evidence that justifies or refutes the intended use of the obtained information as part of the evaluation.
- Make an overall assessment of the inferences or conclusions drawn from the obtained information.

Validity resides not in any instrument or procedure, but in that instrument or procedure's use in generating inferences and conclusions in a particular study. It is incorrect to generalize that an instrument or procedure is valid or not valid. Instead, an evaluator should judge as valid or not valid the inferences or conclusions emanating from a particular use of an instrument or procedure. The key determinant of validity is how fully and dependably the obtained information answers the study's questions. Evaluators should avoid the common mistake of assuming that their intended use of a procedure or instrument is justified because an investigator reported high validity in another study. Instead, evaluators need to validate their inferences or conclusions pursuant to a study's particular questions and based on assessments of the study's procedures, instruments, and obtained information.

Due to feasibility constraints, evaluators often have to employ tools and procedures whose uses in the particular study do not evidence optimal validity. To counter this, they should employ multiple information collection methods to provide checks and balances on possibly weak measures and ensure that the combination of methods effectively addresses all of the study's questions. They should validate the inferences and conclusions resulting from multiple measures individually and in combination to ensure that the obtained information is pertinent, sufficient, and defensible. They should also report weaknesses in the obtained information and, as appropriate, warn audiences to be cautious in using the findings.

Shown in Exhibit 22.6 is a checklist of practical tasks to assist evaluators in meeting Valid Information standard requirements.

Exhibit 22.6 VALID INFORMATION CHECKLIST

Checkpoints

_____ Through communication with the full range of stakeholders, develop a coherent, widely understood set of concepts and terms needed to assess and judge the program within its cultural context.

_____ Ensure—through such means as systematic protocols, training, and calibration—that data collectors competently obtain the needed data.

_____ Document the methodological steps taken to protect validity during data selection, collection, storage, and analysis.

_____ Involve the client, sponsor, and other stakeholders sufficiently to ensure that the scope and depth of interpretations are aligned with their needs and are widely understood.

_____ Investigate and report threats to validity, such as by examining and reporting the merits of alternative explanations of program outcomes, for example.

_____ Assess and report the comprehensiveness, quality, and clarity of the information provided by various data collection procedures in relation to the information needed to address the evaluation's purposes and questions.

_____ Engage program personnel and other stakeholders to check proposed information collection tools and procedures against the evaluation questions.

_____ In choosing or developing information collection tools, pay close attention to the characteristics of the intended respondents (for example, their reading ability, physical disabilities, conflicts of interest, or native language) that might affect the validity of their responses.

_____ Obtain and report validity evidence from other similar studies that used the same evaluation tools.

_____ Follow sound instrument development procedures to minimize biased or confused answers from respondents; for example, in each item of a rating scale, include only a single point to be rated.

_____ Select appropriately qualified information collection personnel; provide them with an orientation, training, support, and supervision; and document their qualifications and performance in the evaluation report's technical appendix (or in a separate technical report).

_____ Carefully plan, monitor, supervise, and document the information collection process.

_____ Document and report significant contextual influences on information collection in the evaluation report's technical appendix (or in a separate technical report).

_____ Report the relevant validity claims and evidence for each evaluation tool and procedure and for all of them in combination in the evaluation report's methods section or technical appendix (or in a separate technical report).

Information Management

The standard states, "Evaluations should employ systematic information collection, review, verification, and storage methods" (Joint Committee, 2011, p. 193).

Systematic information control is an information management process to ensure that an evaluation's information is regularly and carefully checked, made as error-free as possible, and kept secure. There are numerous errors to avoid, which include mistakes in collecting, scoring, coding, recording, organizing, filing, releasing, analyzing, and reporting information. Information might be collected from the wrong respondents. Interviewers might not adhere to the interview protocol. Information coders might not apply coding guidelines correctly. Data files might be misplaced. Unauthorized persons might access filed information. Data might be analyzed inappropriately or incorrectly. There might be clerical errors in the preparation of reports. Results might be reported without needed caveats concerning errors that were discovered but not corrected. Erroneous data might be included in reports. Report findings might be leaked. These are only some of the things that can go wrong in the course of obtaining, processing, storing, and reporting information.

Evaluators should institute safeguards to prevent all such mistakes. Otherwise, members of the audience could be misled to place unwarranted confidence in erroneous information, or the evaluation might become the center of controversy. When mistakes are uncovered belatedly, the evaluation is likely to be discredited and rendered useless. A sound information management process includes systematic orientation and training of evaluation personnel, close supervision of all aspects of the evaluation, and checks for accuracy. It also involves a secure filing system including rules and systematic procedures for accessing, reviewing, and replacing files. Evaluators should maintain control of original information and results and, as appropriate, make copies for use by coders and analysts. Often they should engage persons who supplied information to review the information for accuracy (that is, member checks). It is especially important to verify data entry for accuracy and to proofread data tables and other renderings of evaluative information.

To summarize, Exhibit 22.7 shows a checklist of key ways to meet the requirements of the Information Management standard.

Exhibit 22.7 INFORMATION MANAGEMENT CHECKLIST

Checkpoints

_____ Select information sources and procedures that are most likely to meet the evaluation's needs in regard to accuracy and to be respected by the evaluation's client group.

_____ Ensure that the collection of information is systematic, replicable, adequately free of mistakes, and well documented.

_____ Establish and implement protocols for quality control of the collection, validation, storage, and retrieval of evaluative information.

_____ Document and maintain both the original and the processed versions of obtained information.

_____ Retain the original and analyzed information as long as authorized users need it.

_____ Store the evaluative information in ways that prevent direct and indirect alterations, distortion, destruction, or decay.

Program Evaluation Standards' Key Themes Concerning Information Collection

As seen in this section (and in Chapter 3), the evaluation field has developed a set of strong standards to help evaluators obtain defensible information. We advise evaluators to master and regularly apply professional standards for evaluations. The Joint Committee (1994, 2011) has provided many helpful references accompanying each of its standards. The seven standards highlighted and elaborated here are especially helpful in fostering the collection of sound information. A theme that runs through these standards is the necessity of employing multiple information sources and multiple procedures. These are needed to cover the scope of needed information and to provide checks and balances on individual procedures that, for practical reasons, often cannot be fully validated. Counterbalancing the standards' emphasis on adequate scope of obtained information is the admonition that, for feasibility reasons, evaluators must also set priorities to ensure that the most important information will be collected. Also, most evaluations should be based on obtaining, analyzing, and synthesizing qualitative and quantitative information. Another important theme in the section has been that evaluators should document and report in detail their information collection procedures and the strengths and weaknesses of the obtained information. As already noted, we recommend that evaluators include such information in a technical appendix or separate technical report.

An Information Collection Framework

Table 22.1 offers a framework for planning an evaluation's information collection component. It is intended to help evaluators consider a comprehensive set of potential information needs and a wide range of possibly relevant information collection procedures, and subsequently to

Table 22.1 An Example Framework for Planning an Evaluation's Information Collection Component

	Areas of Information						
Information Collection Procedure	**Program Context and Recipient Needs**	**Program Plan and Competing Approaches**	**Program Activities and Costs**	**Program Reach to Targeted Recipients**	**Program Outcomes**	**Program Sustainability**	**Program Transportability**
Documents	✓	✓	✓	✓	✓	✓	
Literature review		✓					
Interviews	✓		✓	✓	✓	✓	✓
Traveling observers			✓				
Site visits		✓	✓			✓	
Surveys							
Focus groups				✓		✓	✓
Hearings							
Public forums							
Observations			✓				
Case studies					✓		
Goal-free evaluation					✓		
Knowledge tests					✓		
Self-assessments							

relate the two. Along the top of the table are seven areas of information needs drawn from the context, input, process, and product (CIPP) model, which was presented in Chapter 13. Arrayed down the vertical dimension are fourteen techniques that we and other evaluators have found particularly useful in a wide range of evaluations. The check marks in the cells illustrate how an overall information collection plan might be charted and summarized. Such a summary is especially useful to assess the extent to which the plan provides for multiple measures of each area of information and to communicate the overall scheme to the client. The rows with blanks illustrate that not all procedures are necessarily relevant or needed in particular studies.

Table 22.1 is intended for use in conceptualizing an overall master plan of information collection. Subsequently, additional tables can be constructed to elaborate the information collection plan. Evaluators can adapt the table by replacing its horizontal dimension with a timeline to show which procedures will be applied at which times. This analysis helps avoid collecting too much information at any one time and also is an aid to scheduling information collection activities. Table 22.2 is an illustration of this adaptation of Table 22.1.

Tables 22.1 and 22.2 summarize plans at the macro level and are useful especially for communicating with clients. In addition, evaluators require plans at the micro level to guide the specific work of information collection. Each information area should be broken out in terms of specific information needs. Table 22.3 illustrates how this is done in relation to the information area of program outcomes. In this table, the information area has been divided into

Table 22.2 Illustrative Timeline for Applying an Evaluation's Different Information Collection Procedures

	Time Periods in the Evaluation						
Information Collection Procedure	**Period 1 (Start-Up and Context Evaluation)**	**Period 2 (Input Evaluation)**	**Period 3 (Process Evaluation and Cost Analysis)**	**Period 4 (Process and Impact Evaluations)**	**Period 5 (Outcome Evaluation)**	**Period 6 (Sustainability and Transportability Evaluations)**	**Period 7 (Final Report Preparation and Delivery)**
Documents	✓	✓	✓	✓	✓	✓	
Literature review		✓					
Interviews	✓		✓	✓	✓	✓	
Traveling observers			✓	✓			
Site visits		✓	✓			✓	
Surveys							
Focus groups				✓		✓	
Hearings							
Public forums							
Observations			✓	✓			
Case studies					✓		
Goal-free evaluation					✓	✓	
Knowledge tests					✓		
Self-assessments							

Table 22.3 Framework for Planning an Evaluation's Information Collection Procedures

Information Collection Procedure	Intended Effects	Side Effects	Cost-Effectiveness
Documents	✓		✓
Literature review			
Interviews	✓	✓	
Traveling observers			✓
Site visits			
Surveys			
Focus groups			
Hearings			
Public forums			
Observations			
Case studies	✓	✓	
Goal-free evaluation	✓	✓	
Knowledge tests	✓		
Self-assessments			

intended effects, side effects, and cost-effectiveness. Similar derivative tables can be constructed for each area of information.

Useful Methods for Collecting Information

We have recommended that evaluators consider and selectively apply a variety of information collection methods. These include qualitative and quantitative methods, and they are used to obtain information from a wide range of sources. Such sources include existing records and other printed material; relevant publications; and the full range of program stakeholders, including especially the beneficiaries and program personnel, other interested parties, experts with relevant expertise, and the evaluators. In this section we describe and comment on some methods that we have found particularly useful in evaluations but that are not widely discussed in the evaluation literature.

Document Retrieval and Review

As a general practice, it is wise to start the information collection process by identifying and collecting relevant existing information for analysis. The practice of collecting and using such information enhances both the scope of obtained information and efficiency in the information collection process. Working from existing information can produce cost savings for the evaluation.

This practice may also enhance accuracy, as much of the information will have been assessed systematically and edited. Using existing information is considerate to respondents, because an evaluator will not need to ask them to supply information that is already collected, sound, and accessible. However, the use of existing information does not entirely limit evaluators' questioning of stakeholders about that information. Evaluators often need stakeholders to assist in verifying the accuracy of the information, cross-checking areas that may be in conflict, and clearing up ambiguities. In selecting and using existing information, evaluators should remember that it was obtained for purposes other than answering the questions of the evaluation at hand, and should therefore ensure that the information is valid for its intended use in the particular evaluation. As Table 22.2 has shown, the collection and analysis of existing documents continues throughout the evaluation process.

Existing information of potential use in an evaluation may be of many types. Exhibit 22.8 provides a checklist of some of the files and documents that may be relevant in particular studies. For convenience, we have grouped the items of information according to where they are likely to be found. The left-hand column contains information that typically exists outside the program and its home institution, and the right-hand column presents information more likely to be present in the program or its home institution. In the case of an evaluation of a national or state-level program, however, some of the information in the left-hand column might be considered internal information. The main points of the exhibit are that evaluators should consider a broad range of documents and files that may be responsive to evaluative questions of interest, and then should use those that are found to be relevant and valid for the intended use.

Exhibit 22.8 CHECKLIST OF DOCUMENTS AND OTHER INFORMATION OF POTENTIAL USE IN AN EVALUATION

Often External to a Program	Often Internal to a Program
_____ Community demographic information	_____ Statistics on targeted participants
_____ Census reports	_____ Needs assessment reports
_____ Consumer reports	_____ Institutional mission statement
_____ Journal articles	_____ Strategic plan
_____ Almanacs	_____ Curricula
_____ Encyclopedias	_____ Collective bargaining agreement
_____ Magazines	_____ Institutional policies handbook
_____ Laws and statutes	_____ Program proposal
_____ Court records	_____ Program progress reports
_____ Police reports	_____ Program evaluation reports
_____ Real estate records	_____ Minutes of meetings
_____ Chamber of commerce records	_____ Staff résumés
_____ Accreditation standards	_____ Program budgets
_____ State standards	_____ Program financial records
_____ State achievement test reports	_____ Accounting reports
_____ National achievement test reports	_____ Audit reports
_____ Polls	_____ Log of visitors to the program
_____ National survey reports	_____ Correspondence
_____ State survey reports	_____ Local achievement test reports
_____ Local survey reports	_____ School district attendance records
_____ Newspaper articles	_____ School district graduation records
_____ National data sets	_____ School district discipline records
_____ State data sets	_____ Local survey reports
_____ Congressional records	_____ Hospital charts
_____ White house reports	_____ Immunization records
_____ Government department reports	_____ College admission records
_____ Professional society reports	_____ College graduation records
_____ Health department reports	_____ Local data sets
_____ Stock market indexes	_____ Accident reports
_____ Internet sites	_____ Insurance records
_____ Information clearinghouse documents	_____ Publicity releases
_____ Other	_____ Other

The most pertinent existing information is likely to be available at the program site. In identifying and accessing this information, the evaluator should consult the client, and together they should institute safeguards against violating the rights of anyone associated with the information. Other relevant information may be found by conducting searches on the Internet and visiting the local library, newspaper offices, government agencies, social service organizations, and other organizations. The evaluator should plan and budget as required for retrieving and assessing relevant existing information.

Literature Reviews

A special case of retrieving and analyzing existing information is the standard literature review, as typically employed in doctoral dissertation research. Literature reviews have two particular uses in evaluations. First, when planning an evaluation, an evaluator may obtain ideas and instruments by identifying and examining the methods and tools used in similar evaluations. Second, the evaluator can conduct a literature review of evaluation and research reports to assist in answering one or more of the evaluation's substantive questions. Conducting, reporting on, and otherwise using literature reviews lend scholarly credibility to an evaluation and also can save time and resources that would otherwise be devoted to devising instruments or collecting information that would only duplicate previous efforts. Obviously the Internet is a valuable source of information.

Each of the two types of literature review starts with a specific question. For example, in the case of planning an evaluation, one might focus the literature review to determine what procedures and instruments have been used to assess preschool children's immunization needs. In the case of answering a substantive question, one might seek national statistics on the incidence of attention deficit/hyperactivity disorder (ADHD) among first- and second-grade students. This could be part of the evaluation's needs assessment.

To investigate either question, an evaluator might begin by doing an informal exploratory search of the applicable literature, perhaps with the assistance of an expert consultant or a librarian. Subsequently, the evaluator would need to define the search parameters. These could include (1) documents published within a set time period; (2) reports only from doctoral dissertations, specified refereed journals, and externally funded evaluations; and (3) documents containing key words from the question being addressed. The evaluator would next use an appropriate computer search engine to identify documents that meet the search criteria. Subsequently he or she would screen these to cull documents that do not address the question of interest. Next, the evaluator would systematically review the remaining documents to identify pertinent responses to the question of interest and evaluate the quality of those responses. The results from this review would be studied to identify areas of agreement and contradiction. In addition, the evaluator would scrutinize references in the documents carefully and obtain and study additional relevant documents that were not in the original set. He or she could then analyze and synthesize the obtained information and combine it with other information to answer the question of interest.

In the case of the first example just given, the evaluator would use the information to choose or develop methods and instruments for conducting the projected evaluation, paying

special attention to validity and reliability evidence related to identified tools and methods. In regard to the second example, the evaluator would use the literature review results to assist in reporting the national incidence of ADHD in first- and second-grade children. Such information could provide a valuable baseline against which to examine and interpret local statistics on the disorder.

Interviews

Among the most useful evaluation methods is the interview. This procedure enables the evaluator to obtain descriptive and judgmental information from a wide range of persons who have important perspectives on a program, its setting, or its beneficiaries. Interviews may be highly structured and inflexible, as in the case of many telephone interviews; relatively unstructured and exploratory; or quite structured but flexible in their administration. They may be conducted with individuals or groups, face-to-face or over a telephone. All of these variations of interviews can be highly informative. What they have in common is a quest to obtain valuable information for use in understanding and judging a program or other evaluand or to obtain leads for pursuing additional information sources.

Whatever the type of interview, its effectiveness and fairness depend on a number of common factors. The interview should be thought through in advance and well planned in terms of the information being sought. The interview protocol should be drafted as clearly as possible, critiqued by others, pilot-tested, and refined. Interviewees should be selected carefully, though not necessarily to represent a population. Depending on the purpose of the interviews, interviewees may be chosen based on random, purposive, or snowball sampling. When possible, they should be contacted in advance to request their agreement to participate. They should be informed of the evaluation's purpose and the roles of interviewee and interviewer. They should also be informed of the amount of time for the interview. In our experience, most interviews should consume fifty minutes or less. Interviewees should be informed whether their responses will be anonymous or kept in confidence. If not, the evaluator should either obtain written permission to associate responses with the particular respondent or not proceed further with that interviewee. If the evaluator desires to tape-record the interview, he or she should so inform the interviewee and obtain written permission. If the interviewee declines this condition, the evaluator should agree to use paper and pencil to document interviewee responses. As appropriate, evaluators also should consult prospective interviewees about when and where the interview should be conducted. In some of our evaluations, interviewees preferred to be interviewed at home. This approach can have the advantage of observing the interviewee in his or her home environment, but it is also subject to distractions. We have sometimes experienced children running about, telephone interruptions, a television playing in an adjoining room, visitors arriving to observe, and a chain saw roaring just outside the house. In general, we prefer a neutral, quiet site for conducting face-to-face interviews; nevertheless, there can be good reasons to conduct an interview in an interviewee's home or some other setting.

When the interviewee agrees to participate, the interview should be scheduled at a time that is convenient for him or her, which in some cases could be immediately. Prior to an interview scheduled for a future time, it is prudent to telephone the interviewee or send him

or her a note with a reminder of the approaching interview, as many doctors and dentists do to help prevent no-shows.

It is important to establish rapport with the interviewee at the outset. The interviewer might review the interview's purpose, indicate that the interviewee is deemed to possess a valuable perspective on the subject program or other evaluand, review the prior agreements under which the interview will be conducted, state appreciation for the interviewee's participation, reiterate how much time the interview will require, and invite and respond to any questions from the interviewee. Usually this initial orientation and exchange are sufficient to establish rapport for a productive interview.

As already mentioned, when the interviewer proceeds to conduct the interview, there should be some means of recording the interviewee's responses. The interviewer might check multiple-choice options on the interview protocol as the interview proceeds (especially in the case of a telephone interview), write out the interviewee's responses, or tape-record the session if prior permission has been obtained from the interviewee. When feasible, it can be especially productive to have two members of an interview team present. One conducts the interview, and the other keeps notes and, as appropriate, interjects with follow-up questions, especially when responses were not clear. In successive interviews, the interviewers can intermittently exchange interviewing and note-taking roles. As an interview proceeds, the interviewer (or interviewers) should, as needed, ask the interviewee to clarify or elaborate on unclear or incomplete responses. At the end of the interview, it is a good idea to invite the interviewee to add any information she or he views as important. Finally, the interviewer should thank the interviewee for her or his valuable contribution to the study.

Following an interview, it is important to review the written record of the responses as soon as possible, while the memory of the session is fresh. Because a written record of the interview is likely to be cryptic, this is the time to add details that one remembers but did not write down. Clearly this point has implications for scheduling when multiple interviews are involved. If possible, the evaluator should schedule time following each interview to scrutinize the results and flesh out the record. This activity may not be necessary if the full interview session was tape-recorded. In that event, the tape should be transcribed for review and analysis. The preceding discussion is intended to apply generally to all types of interviews. However, different types have some unique requirements worth mentioning:

- *Telephone interviews*. Typically multiple interviewers conduct the interviews over the telephone and code responses as they are received. The interviews must be administered according to a standard protocol, which needs to be scripted carefully so that all interviewers will obtain comparable data that can be aggregated and analyzed. The interviewers should be thoroughly trained, calibrated, and supervised. Usually the interviewees in telephone interviews are selected randomly or systematically to ensure that they are representative of a population of interest. An advantage of tightly scripted telephone interviews is that they are quite amenable to scoring, aggregation, and statistical analysis. A disadvantage is that they are not sufficiently open ended and flexible to obtain in-depth information that capitalizes on the idiosyncratic insights of different respondents.

- *Semistructured, flexible interviews.* These interviews are much harder to summarize, aggregate, and subject to statistical analysis than telephone interviews. Nevertheless, they can yield invaluable qualitative information, and their results are amenable to qualitative analysis and identification of important themes. In using this approach, the evaluator usually wants to gain insights into the strengths and weaknesses of a program from a wide range of perspectives. At the outset, he or she may have in mind a set of questions and persons who could answer them. However, too much structure might prevent the evaluator from obtaining a rich set of insights from parties other than those on the initial list of interviewees. Such respondents might identify key issues in a program that the evaluator has not thought to investigate previously. Thus, the evaluator might start by contacting a few known stakeholders. In the initial interview with each respondent, the evaluator might ask the respondent to identify and discuss what he or she considers to be the most important issues in the subject program. Near the end of the session, the interviewer would ask the interviewee to identify anyone who might have additional insights into the issues discussed. Subsequently, the evaluator would follow up the obtained leads and contact the identified persons. These new contacts also would be asked to identify issues and other persons who could shed light on them. From interview to interview, the evaluator would review and keep a record of what is being learned. At the end of the process, he or she would have a rich set of information to examine, analyze, and synthesize. In this approach, the evaluator employs snowball sampling to choose interviewees.
- *Structured but flexibly administered interviews.* In this approach, the evaluator prepares a structured set of questions, which if possible will fit on a single sheet of paper. The interviewer might provide the interviewee with a copy of the interview questions as an agenda and a heuristic. The interviewer then starts the process by asking the first question. As the interviewee responds, he or she may expand beyond stated questions and begin to answer other questions further down the list. As long as responses are relevant to the established questions, the interviewer allows the respondent to move through them in any sequence that helps him or her get the message across. In this approach, the interviewer is concerned with obtaining in-depth, coherent responses to all the questions, but is not concerned about the sequence of responses. In some cases, the evaluator might not employ a printed list of the questions, instead holding the set of questions in her or his mind. Here, a skillful interviewer asks a question to start the interview and then engages in a free-flowing discussion with the interviewee. The evaluator mentally keeps track of the extent to which all the questions are being addressed in whatever sequence and steers the discussion to make sure that all questions are answered. In this approach, it is desirable that the session be tape-recorded. Alternatively, if tape-recording is not agreed to in an interview of this type, we have found it advantageous to have two interviewers in the interview session: one to administer the questions and the other to keep detailed notes and ask follow-up questions.

Interviewing is one of the most pervasive, adaptable, and valuable procedures for gathering evaluative information. Although there are alternative acceptable approaches to interviewing,

all forms of the procedure should be applied with careful planning and rigor. For additional information about interviewing methods, see Gubrium and Holstein (2001), Kvale (1996), and Seidman (2006).

Focus Groups

A variation of the interview is the focus group procedure, a group interview approach developed in the consumer research field and widely used during the 1950s and 1960s. The technique has since been adapted and applied for several different purposes. After (or preceding) elections, focus groups often are seen on television, with a moderator engaging voters or likely voters to discuss election issues or results. Evaluators frequently use the technique to obtain and analyze the views of stakeholders concerning the merit and worth of a subject program or pertaining to given evaluation questions.

Originally researchers employed this technique to engage a sample of consumers to judge a consumer product or service. As part of the usual procedure, researchers recruited about a dozen consumers—usually in a typical community—and interviewed them as a group to hear their individual and collective judgments of a product or service they had tried. The interviews would last up to two hours and focus on questions of particular interest to those who developed and marketed the product or service. In starting the session, the moderator would stipulate that each person's perception was important, that there were no right or wrong answers, that participants should feel free to agree or disagree, and that they should probe each other's responses in the interest of providing in-depth understanding and revealing key areas of agreement and disagreement. Especially, the focus group members were asked to be advocates for the potential consumers of the subject product or service. Accordingly, they were expected to send a message to developers concerning what people like them needed and expected of the product or service and what they saw as good and bad about the one they had tried. The moderator's responsibilities were to draw all panelists into the ensuing discussion of each question, keep the interview moving, ask follow-up questions to promote clear and in-depth responses, and prevent any one member from dominating the discussion. An observer would make a written record of the interchange, and the session would probably be tape-recorded or videotaped. The investigator subsequently would analyze the focus group record to identify areas of agreement and disagreement and discern important themes. The target audience of developers and marketers would use the focus group findings along with other information to make decisions related to modifying, packaging, advertising, and selling the subject product or service.

The evaluation field adapted and began using the focus group technique in the 1970s, when evaluators had begun to expand their methods into the realm of qualitative approaches. In the early stages of this movement, evaluators mainly borrowed qualitative methods from other fields, including jurisprudence, sociology, psychology, ethnography, and consumer research. In the focus group procedure, evaluators found a ready-made tool for systematically obtaining multiple perspectives on given evaluation questions. This technique provided some of the benefits found in individual interviews and provided insights based on the interplay among

multiple respondents during a single session. It is noteworthy that an evaluation might employ multiple focus groups—for example, one for staff, one for beneficiaries, one for policy board members, and one for subject matter experts. Generally the membership of each focus group should be homogeneous.

Evaluators have tended to stay true to the original intent and procedures of focus groups but also have made some changes in the technique. Drawing from our experience in using the technique, we offer the following recommendations for selecting and engaging focus groups to help evaluate programs:

1. Determine a homogeneous class of potential members of the focus group.
2. Stipulate the issue that the focus group will address.
3. Determine a sequence of questions designed to move the group's discussion toward the issue of interest.
4. Select seven to ten members to participate; they should share a common perspective, such as that of beneficiary, but otherwise should reflect the diversity of the larger group that shares their perspective in regard to gender, age, education, ethnicity, and other matters.
5. Allot one to two hours for the session.
6. Hold the meeting in a setting that is comfortable, free from distractions, and conducive to discussion. It might be a round table or a circle of easy chairs in a quiet room.
7. Make a record of the group discussion using a tape recorder, video recorder, or written notes.
8. Provide all members of the group with a common orientation to the meeting's objectives, the agenda of questions, relevant background information, meeting structure, the role they will play, and the time allotted for the session.
9. Stress that each person's perception is important, that participants should feel free to agree or disagree, and that they should probe each other's responses in the interest of providing in-depth understanding and revealing key differences of opinion.
10. Conduct the session within a permissive, nonthreatening atmosphere.
11. Skillfully keep the discussion focused on the meeting's objectives; move the conversation through the agenda of questions; ask follow-up questions to promote clear, in-depth responses; and prevent any one member from dominating.
12. Ensure that all group members are given the opportunity to participate and are encouraged to do so.
13. In concluding the session, invite each member to state what he or she judges to be the one or two most important points made during the session.
14. Thank everyone for participating, and adjourn the meeting.
15. Following the meeting, prepare a transcript of the exchange from meeting notes or recordings, for example.
16. Analyze the focus group record to identify areas of agreement and disagreement and to discern important themes.

Traveling Observers and Resident Researchers

The traveling observer (TO) technique, developed at the Western Michigan University Evaluation Center, involves the training and assignment of a field researcher to conduct preliminary investigations in advance of subsequent primary evaluations by a panel of experts. Typically, the TO travels from site to site to contact and develop rapport with data providers, collect preliminary information, and work out a plan for a follow-up site visit team. The TO next provides the site visit team with orientation and training prior to its site visits and may support the team during its on-site investigations. Usually the TO is a relatively junior investigator, with members of the follow-up team being more senior. Often the TO spends considerably more days in gathering preliminary data than the follow-up site visit team will spend. The TO usually is compensated at a considerably lower rate than are members of the site visit team. Thus, an advantage of the technique is that it saves money for a sizable part of the needed field research.

A variation of the technique is the resident researcher technique, which essentially is the application of the TO approach at a single site. (The TO technique is described more fully in Chapter 13.)

Advocate Teams Technique

The advocate teams technique was developed at the Evaluation Center when it was located at The Ohio State University. It was created and applied in 1969 to help the Texas-based Southwest Regional Educational Laboratory identify and assess alternative strategies for serving the acute education needs of migrant children. This technique was created because the methodology of evaluation lacked procedures for identifying and assessing competing strategies for addressing high-priority needs and problems.

The advocate teams technique has five main steps. The first is to stipulate a target group of beneficiaries and identify objectives to be achieved in meeting this group's assessed needs. The second step is to create alternative strategies for achieving the stipulated objectives. The evaluator establishes two or more advocate teams, and each team is provided with the subject objectives, pertinent needs assessment information, criteria for assessing possible program strategies, and a structure for writing up a proposed program strategy. Next, each advocate team studies the needs assessment data and pertinent literature, brainstorms toward inventing an appropriate program strategy, and writes up its proposed program strategy. In the fourth step, an independent panel evaluates the advocate teams' proposals against the predetermined criteria and ranks them on overall merit. The client then chooses a strategy for implementation or may assign a convergence team to merge the best features of the competing plans into a hybrid plan. The technique is keyed directly to a decision-making group's desire for creative solutions to high-priority needs and problems. (Additional information about this technique appears in Chapter 13.)

Additional Techniques

In addition to the techniques reviewed here, we recommend that evaluators be resourceful in searching broadly for techniques that will address the information needs of their studies

effectively. Some of these techniques, or approaches, are discussed in other chapters in this book, including the Success Case Method (Chapter 6), case study evaluation (Chapter 12), and goal-free evaluation (Chapter 14). Others, such as questionnaires and rating scales, data mining, needs assessment, visual methodologies, cost analysis, and ethnography, are treated in a wide range of research and evaluation methodology textbooks (for example, Bickman & Rog, 2009; Davidson, 2005; Fitzpatrick, Sanders, & Worthen, 2011; Margolis & Pauwels, 2011; Rossi, Lipsey, & Freeman, 2004; Wiersma & Jurs, 2009).

Summary

The collection of sound information is essential to the success of any program evaluation. Evaluators must obtain a sufficient range and depth of appropriate and reliable information if they are to reach valid conclusions about a program's merit and worth. The Joint Committee (1994, 2011) has provided authoritative, useful advice for carrying out sound processes of information collection, and evaluators are advised to master and regularly apply the committee's program evaluation standards. Use of these standards can be helpful in determining what information to collect, upholding the rights of human subjects, studying a program's context, fully describing program operations, using appropriate sampling methods, checking and enhancing the reliability of evidence, validating instruments, and maintaining the integrity of obtained information. Because of the practical constraints in almost all program evaluations, evaluators often have to apply less-than-perfect instruments and procedures. They are advised to employ multiple methods to allow for cross-checks in their search for consistent findings and also to report the limitations of the information they obtain. Evaluators should plan their data collection efforts to fulfill the study's information requirements and also to uphold the rights of respondents and not impose undue burdens on program participants. Finally, the research and evaluation fields have produced a wide range of information collection techniques, and evaluators are advised to make good, selective use of the available information collection tools and strategies—both quantitative and qualitative.

REVIEW QUESTIONS

1. The validity of evaluative conclusions depends heavily on the adequacy of the information used to reach those conclusions. In general, what requirements should be met by the information that an evaluator uses to judge a program's merit and worth?
2. This chapter has discussed the relevance of certain Joint Committee standards for guiding and assessing the collection of information in an evaluation. Explain the relevance of each of the following standards to the information collection task: Relevant Information, Human Rights and Respect, Explicit Program and Context Descriptions, Defensible Information Sources, Reliable Information, Valid Information, and Information Management.

3. Why is it misleading or even incorrect to state that a given evaluation instrument is valid?
4. Distinguish between the terms *valid information* and *reliable information,* and explain why validity is dependent on reliability.
5. What is the basis of the recommendation that evaluators often should employ multiple methods of information collection?
6. List steps you would follow to validate a data collection instrument to be used in a particular evaluation.
7. What are the main benefits and also some of the hazards of obtaining and studying existing information as a partial basis for judging a program?
8. What is the role of the literature review in program evaluations?
9. What is the traveling observer technique? How is it used in program evaluations? What are some advantages of employing this technique in a program evaluation?
10. What is the focus group procedure, and what are its uses in program evaluations?

Group Exercises

Exercise 1

Define each of the following techniques, and then develop an example of how you might beneficially apply each of the techniques in an evaluation: simple random sampling, stratified random sampling, purposive sampling, snowball sampling, and studying an entire population.

Exercise 2

Develop an illustrative case showing the relevance of and steps involved in applying the advocate teams technique.

Exercise 3

Develop a checklist of points to observe in planning and conducting sound interviews of a program's beneficiaries.

Exercise 4

A small county hospital had been in the news for all the wrong reasons: two forced resignations of chief administrators within the past three years, high staff turnover based on what appear to be legitimate grievances, an evident shortage of funds to meet some basic medical requirements, and a series of minor scandals involving medical and nursing staff. This sad chronicle of events was capped by the death of two patients resulting from salmonella contamination in the hospital's kitchen. County taxpayers, the hospital's stakeholders, could endure these

catastrophes no further. They called a special general meeting that forced the hospital board to initiate immediate evaluation of the institution's procedures, finances, culture (including its effectiveness as a health care provider), and future viability for the community. The board issued a general request for proposal (RFP), part of which required implementation of the focus group procedure to obtain stakeholder input.

Your group has decided to respond to this RFP. Your assignment in this exercise is to answer convincingly the following questions concerning the focus group component:

1. What perspectives would be important to seek out in choosing the members of the focus group? Justify your response.
2. How many persons would you include in the group, and why?
3. In selecting the members of the focus group, what sampling procedure would you use, and why?
4. What roles would evaluators need to take on to conduct the focus group? What responsibilities would you assign to the evaluator in each role? Justify your answers.
5. Outline the main questions to be addressed in the focus group session. Justify your response.
6. To what extent would you follow a strict agenda, as opposed to allowing a totally free-flowing discussion? Justify your answer.
7. Outline an agenda for this session. Justify your plan.
8. How much time would you allow for conducting the session? Justify your response.
9. Outline the contents of your projected report of the focus group session, and justify your plan for this part of the overall evaluation report.
10. List criteria for assessing the focus group segment of the proposed evaluation, and justify these criteria.

Conclude this group exercise by having each group member briefly state a known situation in which the focus group procedure would be appropriate for use in collecting evaluative information.

Suggested Supplemental Readings

Bamberger, M., Rugh, J., & Mabry, L. (2012). *RealWorld evaluation: Working under budget, time, data, and political constraints* (2nd ed.). Thousand Oaks, CA: Sage.

Bickman, L., & Rog, D. J. (2009). *Handbook of applied social research methods* (2nd ed.). Thousand Oaks, CA: Sage.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Davidson, E. J. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation*. Thousand Oaks, CA: Sage.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* (3rd ed.). Hoboken, NJ: Wiley.

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2011). *Program evaluation: Alternative approaches and practical guidelines* (4th ed.). Upper Saddle River, NJ: Pearson.

Gubrium, J. F., & Holstein, J. A. (Eds.). (2001). *Handbook of interview research: Context and method*. Thousand Oaks, CA: Sage.

Henry, G. T. (1990). *Practical sampling*. Applied Social Research Methods Series, Vol. 31. Thousand Oaks, CA: Sage.

Kish, L. (1965). *Survey sampling*. Hoboken, NJ: Wiley.

Kvale, S. (1996). *InterViews: An introduction to qualitative research interviewing*. Thousand Oaks, CA: Sage.

Lohr, S. L. (2010). *Sampling: Design and analysis* (2nd ed.). Belmont, CA: Thompson.

Margolis, E., & Pauwels, L. (Eds.). (2011). *The Sage handbook of visual research methods*. Thousand Oaks, CA: Sage.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.

Scheaffer, R. L., Mendenhall, W., III, Ott, R. L., & Gerow, K. G. (2012). *Elementary survey sampling* (7th ed.). Belmont, CA: Thompson.

Seidman, I. (2006). *Interviewing as qualitative research: A guide for researchers in education and the social sciences* (3rd ed.). New York, NY: Teachers College Press.

Wiersma, W., & Jurs, S. G. (2009). *Research methods in education: An introduction* (9th ed.). Boston, MA: Pearson.

CHAPTER 23

ANALYZING AND SYNTHESIZING INFORMATION

The objectives of collecting information in an evaluation are to provide an evidentiary basis for answering priority questions and to allow the evaluator ultimately to judge the program or other object of interest. To finalize an evaluation, the evaluator needs to proceed beyond the collection of information and work through the subsequent processes of analyzing and synthesizing the obtained quantitative and qualitative information and ultimately reporting and supporting use of the findings.

This chapter presents practical advice on the analysis and synthesis tasks. By analysis, we mean identifying and assessing the constituent elements of each set of obtained information and their interrelationships to clarify the information's dependability and meaning for answering particular questions. By synthesis, we mean combining analysis findings across information collection procedures and devices to discern their validity and aggregate meaning for answering the audience's questions and judging the value of the object of interest. For purposes of explanation and illustration, analysis and synthesis are presented in this chapter as independent stages in the evaluation process. In reality, these processes are dependent on and part of the other evaluation processes—design and preparation, collection of information, and reporting—and should be considered and planned for throughout the entire evaluation process. We have organized our discussion in this chapter in terms of quantitative analysis, qualitative analysis, and justified conclusions. We have grounded much of this presentation in writings on analysis and synthesis by the Joint Committee on Standards for Educational Evaluation (1994, 2011). Because we have judged the1994 edition of the Joint Committee's *Program Evaluation Standards* to

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- The rationale for conclusion-oriented evaluation
- A general orientation to and principles for analyzing and synthesizing information
- Definitions and examples of descriptive, relational, and causal questions
- The quantitative analysis process and a range of associated techniques
- Software packages for analyzing data
- Definitions of and threats to internal and external validity
- The concepts of statistical hypothesis testing, Type I and Type II errors, and confidence intervals
- Determining practical significance through calculations of effect sizes
- The role, nature, process, and procedures of qualitative analysis
- The process of and special procedures for synthesizing quantitative and qualitative information
- Bottom-line steps in producing justified conclusions

include valuable, focused guidance on analysis and synthesis that is not as well covered in the 2011 edition, we have elected in this chapter to draw ideas from both editions, in the interest of offering concrete, practical advice.

General Orientation to Analyzing and Synthesizing Information

We believe that evaluations should include divergent as well as convergent stages and typically, but not always, culminate in bottom-line conclusions. Thus, most evaluations should not end on a note of multiple or conflicting answers and interpretations. Summative evaluations that leave conclusions to the eye of the beholder mainly add to confusion and controversy. Open-ended evaluation findings often baffle audiences and leave clients and sponsors wondering why they commissioned and funded the summative evaluation in the first place. Our position is that evaluators initially should search out multiple findings and interpretations (in the evaluation's divergent stage) but subsequently work toward delivering the best answers they can find (in the convergent stage). They should justify their conclusions by documenting the assumptions, rules, and procedures used to analyze and synthesize information. In addition, they should buttress their conclusions with appropriate caveats concerning any deficiencies in the obtained information and possible disagreements about value bases for interpretation. Exceptions to summative evaluations are studies commissioned mainly to provide ongoing formative feedback that are not necessarily expected to result in a final summative evaluation report. In regard to summative evaluations that end by noting that conclusions are open to the readers' interpretations and that they might justifiably judge the assessed program as either good or bad, we identify with an expression attributed to President Harry S. Truman. It went something like this: "I am tired of hearing economists conclude that, on the one hand, the economic outlook is such and such, but, on the other hand, it is very different." Truman reportedly commented further that he was seeking a one-handed economist—one with willingness and competence to determine and commit to a particular interpretation of the available evidence. In regard to an evaluator's risk of possibly making a wrong interpretation, we think another Truman saying has relevance: "If you can't stand the heat, get out of the kitchen." Evaluators who seek and present firm conclusions often face opposition and criticism and sometimes are wrong. However, if they ground their conclusions in systematic analysis and synthesis of a sufficient set of appropriate evidence and report appropriate caveats, we think they will be correct far more often than those who fail to practice systematic, conclusion-oriented evaluation and will be instrumental in helping their audiences make sound decisions and improve programs. Systematic evaluation is not and never will be an exact science, but it is an invaluable guide to progress. Evaluators who are steadfastly afraid of being wrong and consequently equivocate or exercise undue caution

probably should seek other worthwhile work that may involve less ambiguity and risk-taking, such as bookkeeping, proofreading, watch repair, or financial auditing.

Principles for Analyzing and Synthesizing Information

The tradition of evaluation is focused heavily on relatively simple methods of analysis, especially descriptive statistics and tests of statistical significance. Such techniques have been employed to characterize groups and their program-related experiences and outcomes, examine and judge the significance of changes in various indexes over time, contrast and judge the significance of the outcomes of competing programs, identify and assess relationships between variables, and extrapolate findings to predict future outcomes. Beyond statistical analysis, methodologists have advanced procedures for qualitative analysis and final synthesis of findings. Moreover, the art and science of combining quantitative and qualitative methods of data gathering, analyzing, and synthesizing have progressed remarkably in the past two decades. Mixed-method evaluation designs are increasingly employed rather than distinctive quantitative or qualitative designs. In this book we have consistently advocated a broad range of methods to develop a foundation of factual evidence to begin responding to clients' questions. Such procedures are the essence of the divergent phase. Using mixed-method approaches, evaluators may well give greater emphasis to quantitative rather than qualitative procedures, or vice versa, depending on the kind and quality of information that will give validity to responses to evaluation questions, and ultimately to reporting and decision making.

Qualitative analysis techniques are needed to mine and interpret the meaning of such information as testimony, interviews, news accounts, and photographic records. Synthesis techniques are required to converge information from a wide range of quantitative and qualitative analyses into bottom-line judgments; these techniques include the synthesis of facts and values (fact-value synthesis) as well as the synthesis of multiple values (value-value synthesis; see also Coryn, 2007a; Davidson, 2005; Scriven, 2007). Evaluators need to develop facility in selecting and employing procedures for quantitative and qualitative analysis and synthesis to answer the questions of their audiences and reach defensible conclusions about the value of programs or other objects of evaluations. Apt summaries of analysis and synthesis concepts are found in the Joint Committee's 1994 Qualitative Analysis and Quantitative Analysis standards, and in the 2011 Sound Designs and Analyses and Justified Conclusions and Decisions standards. These standards provide a good foundation for discussing the principles and procedures of analysis and synthesis.

The Sound Designs and Analyses standard states: "Evaluations should employ technically adequate designs and analysis that are appropriate for the evaluation purposes" (Joint Committee, 2011, p. 201). The checklist in Exhibit 23.1 identifies actions of use in meeting the requirements of this standard.

Exhibit 23.1 SOUND DESIGNS AND ANALYSES CHECKLIST

Checkpoints

_____ Create or select a logical framework that provides a sound basis for studying the subject program, answering the evaluation's questions, and judging the program and its components.

_____ Plan to access pertinent information sources and to collect relevant, high-quality quantitative and qualitative information in order to answer the evaluation's questions and judge the program's value.

_____ Delineate the many specific details required to collect, analyze, and report the needed information.

_____ Develop specific plans for analyzing obtained information, including clarifying needed assumptions, checking and correcting data and information, aggregating data, and checking for statistical significance of observed changes or differences in program recipients' performance.

_____ Buttress the conceptual framework and technical evaluation design with concrete plans for staffing, funding, scheduling, documenting, and metaevaluating the evaluation work.

_____ Plan specific procedures to avert and check for threats to reaching defensible conclusions, including analysis of factors of contextual complexity, examination of the sufficiency and validity of obtained information, checking on the plausibility of assumptions underlying the evaluation design, and assessment of the plausibility of alternative interpretations and conclusions.

Analysis of Quantitative Information

The Quantitative Analysis standard, from the Joint Committee's 1994 edition of *The Program Evaluation Standards,* states, "Quantitative information in an evaluation should be appropriately and systematically analyzed so that evaluation questions are effectively answered" (p. 165).

Evaluations may encompass a wide range of quantitative information. Examples include age, height, and weight; duration, funding level, expense reports, and ratings of the subject program or other object of the evaluation; and indicators of program outcomes, such as blood pressure readings, weight gain or loss, scores on attitude inventories, number of school years completed, achievement test scores, annual income, and number of traffic violations. When such data are involved, evaluators should employ systematic, rigorous, relevant, and appropriate methods of quantitative analysis. Evaluators should keep in mind, however, that not all evaluations require quantitative analysis.

Most quantitative analyses are used to investigate one or more of three general types of questions:

- *Descriptive questions*. Descriptive questions are the most rudimentary questions that evaluators seek to answer. Answering such questions involves the simple account of a set of observations concerning a set of variables of interest. Few questions of interest to evaluators are exclusively descriptive.
- *Relational questions*. Relational questions are among the most common questions of concern to evaluators. They involve basic assumptions, such as whether a relationship between two or more phenomena exists at all. More often than not such questions are concerned not only with whether a relationship exists between two or more variables but more specifically with the direction and magnitude of covariation.
- *Causal questions*. Causal questions are concerned with whether or how one or more independent variables affect one or more dependent variables. Causal questions can be relatively simple (causal description) or more sophisticated (causal explanation). In general, descriptive causal questions are those with which evaluators inquire as to whether consequences attributable to varying an independent variable can be established, whereas questions about causal explanation are those with which evaluators seek to identify the mechanisms through which, and the conditions under which, causal relationships hold.

Put simply, to describe involves representing or giving an account of something. An evaluator might simply be interested in describing how a state's justice system handles juveniles who commit violent crimes. Are they incarcerated? Are they sentenced to public service? Or are they handled in some other way? An answer to any one of these questions would constitute a description. Although descriptive research sometimes is dismissed as overly simplistic, such inquiry is fundamental to sound evaluations and sometimes can inform public policy decisions, it has added immeasurably to basic knowledge claims, and it often forms the basis for investigating relational and causal questions. For example, descriptive investigations into juveniles and the justice system might raise questions about who is sentenced and to what degree.

The same evaluator in the example just given might then try to determine whether incarceration or public service is related to future behaviors of violent juvenile offenders—for example, by investigating whether there is a relationship between incarceration and the likelihood of committing violent crimes after release, and if such a relationship exists, determining the direction and magnitude of that relationship. That is, does a relationship between sentencing and future violent acts exist? Do the two vary together? Here, the evaluator simply is interested only in establishing the existence of a relationship rather than inferring that the observed relationship is causal, which requires meeting additional assumptions.

Evaluators also sometimes attempt to explain some aspect of human action and interaction, and the social world, through their research. To explain is to give the reason for or cause of a relationship between two or more variables. In this case, the same evaluator might seek to

explain why some juveniles are more likely to commit violent crimes than others. Such explanations can be very general (causal description) or very specific (causal explanation). One general hypothesis might be that juveniles commit violent crimes because their parents hit, slapped, or spanked them as children. Or, more specifically, the evaluator might hypothesize that, through a complex process of social learning, juveniles whose parents (or other caretakers) model violent behaviors first internalize the observed behaviors, then themselves perform the same behaviors, and, through a contingency process of reinforcement, commit increasingly violent acts in adolescence. Continuing the example, then, a simple causal description (examining the whole rather than its constituent parts) might be that juvenile offenders who are sentenced to public service are less likely than those sentenced to incarceration to commit future acts of violence. In this case, the evaluator is interested only in whether sentencing to public service can be causally associated with decreased future acts of violent crime, as contrasted with the results of sentencing to incarceration. That is, does one condition cause another? The evaluator may also inquire into more specific explanations for the causal effect. For instance, does public service cause changes in empathy toward others, which in turn reduces the likelihood of committing a violent crime? Here the question is whether the relationship between public service and the reduced incidence of violent crime is explained by empathy. Evaluators studying such questions seek to determine how rather than when effects will occur by accounting for the relationship between two variables using one or more additional variables.

The general concepts of causal description and causal explanation (which also apply to relational types of questions and provide information that simply describes relationships) are conceptually illustrated in Figure 23.1 using the random variables *X, Y,* and *Z*. In 23.1 A, the effect of variable *X* on variable *Z* is direct ($X \rightarrow Z$), whereas in 23.1 B, the effect of *X* on *Z* is mediated through *Y* ($X \rightarrow Y \rightarrow Z$). In 23.1 C, the effect of *X* on *Z* is both direct ($X \rightarrow Z$) and indirect ($X \rightarrow Y \rightarrow Z$). To illustrate, in case 23.1 A, *Z* is assumed to covary with *X*. That is, to what degree (and in what direction) does incarceration or public service (*X*) predict, or account for, the likelihood of committing violent crimes after release (*Z*)? In the second case, 23.1 B, does sentencing to public service versus incarceration (*X*) cause changes in empathy (*Y*) toward others, which in turn reduces the likelihood of committing violent crimes (*Z*)? In the third case, 23.1 C, does sentencing to public service versus incarceration (*X*) directly cause changes in empathy (*Y*) toward others, which in turn (indirectly through *Y*) reduces the likelihood of committing violent crimes (*Z*), or does *X* directly cause *Z* in the (presumed) absence or presence of *Y*?

As shown in Figure 23.2, such relationships also can be described in terms of whether they moderate (23.2 A) and/or mediate (23.2 B) one another (Baron & Kenny, 1986). In 23.2 A, the effect of *X* on *Z* is moderated by A. That is, *Z* (ignoring *Y* in 23.2 B) varies as a function of the A × B interaction. Here, if A is gender (male or female), the effect (*Z*) differs over different levels of A (gender). That is, the likelihood of recidivism (reincarceration) is different over different levels of gender—*Z* varies over A. Hypothetically, then, the effect of *Z* could be greater (or less) for females (one level of A) than for males (the other level of A). In the mediator model (23.2 B), if empathy (*Y*) explains recidivism (*Z*), then the indirect effect from *X* to *Y* (path A) through *Y* to *Z* (path B), combined, should be greater than zero, whereas path C ($X \rightarrow Z$) should

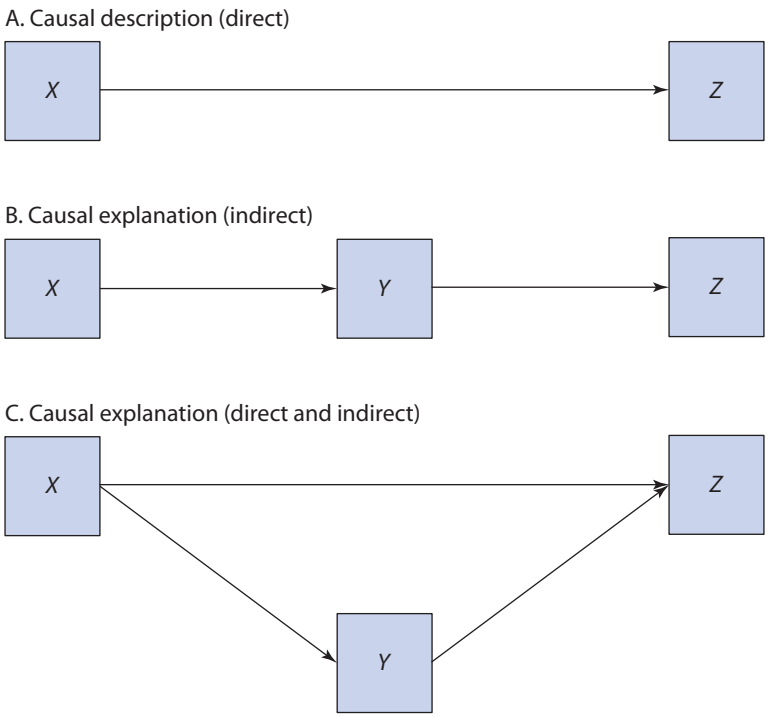

Figure 23.1 Conceptual Illustration of Causal Description and Causal Explanation

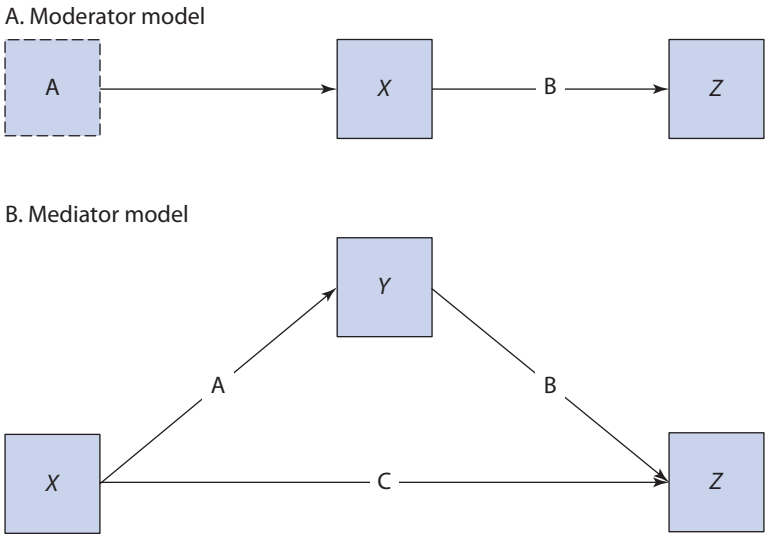

Figure 23.2 Conceptual Illustration of Moderating and Mediating Relationships

statistically equal zero if the effect was truly through *Y*. Both are causal explanations given that they explain causal relationships between one or more independent variables and one or more dependent variables. Not all evaluations require causal explanation, however, and Scriven (1994c) has argued strongly that the role of evaluators is only to determine whether programs work, not to explain how they work.

Notably, most causal reasoning in the social sciences, and evaluation, is probabilistic rather than deterministic. In other words, reasoning in the social sciences and evaluation is not based on general laws or principles—such as the law of gravitational fields used in many of the physical sciences that describes the relationship between force and mass and that has been used in classical mechanics as well as quantum physics. Returning to our example, an evaluator might reasonably assume, based on previous studies, that being female decreases the probability of recidivism following sentencing to public service instead of incarceration, rather than positing that being female always and invariably decreases recidivism following sentencing to public service. Evaluators typically deal in statistical probabilities concerning the likelihood of predictable outcomes rather than in absolutes (such as, an apple always falls to the ground once it drops from a tree due to gravitational forces).

Quantitative Analysis Techniques

Evaluators may choose from a wide range of analysis techniques to examine and interpret quantitative information. These techniques include the following, among others: frequency counts; percentages; histograms; pie charts; trend lines; means and medians; variances and standard deviations; correlations; coefficients of contingency; multiple regression; *t*-tests; chi-square tests; tests of concordance; analysis of variance; multiple analysis of variance; analysis of covariance; a posteriori significance tests; Delphi techniques and *Q*-sorts; gain score analysis; value-added analysis; cost analysis; trend analysis; time-series analysis; pattern analysis; cluster analysis; effect parameter analysis; factor analysis; hierarchical linear modeling, structural equation modeling; discriminant function analysis; concept mapping; multidimensional scaling; meta-analysis; and norm-referenced, criterion-referenced, objective-referenced, and domain-referenced approaches to analyzing achievement test or other scores. Information on such techniques is readily available in a wide range of textbooks on statistics and research methods. Some examples are Goldstein (1987); Hinkle, Wiersma, and Jurs (2003); Hopkins and Glass (1978); Jaeger (1990); Kerlinger (1986); Raudenbush and Bryk (2002); Thompson (2006); Wiersma and Jurs (2005); and Winer (1962). Applications of complex quantitative analysis techniques are facilitated by the use of computers and applicable software. Among the many available statistical packages are Mplus, R, Stata, Statistical Analysis System (SAS), and Statistical Package for the Social Sciences (SPSS). Table 23.1 contains a basic comparison of the major features of these statistical software packages. In the table, the plus signs (from one to five) represent the strengths of each statistical software package relative to the others included in the table. The versatility feature refers to the ability of the statistical software package to perform different types of analysis.

Evaluators must not merely apply their favorite technique and should not allow familiarity with certain techniques and easy access to certain computer programs to dictate the analysis

Table 23.1 Comparison of Statistical Software Packages

Feature	Mplus	R	Stata	SAS	SPSS
Cost for the base package	$695	Free	$1,195	$8,500	$5,120
Platforms	Windows, Mac, or Linux	Windows, Mac, Unix, or Linux	Windows, Mac, or Unix	Windows or Linux	Windows, Mac, or Unix
Versatility	+++	+++++	++++	+++++	+++
Ease of use	++	++	++	++	++++
Data visualization	+	++++	++	++++	++
Technical support	++	++	++	+++	++

process. In approaching a data set, evaluators should consider what the intended audience wants and needs to learn from the data and then choose analysis methods that will best address the focal questions and fit the data's characteristics. The selected methods may involve qualitative as well as quantitative techniques, and sometimes should include only one type or the other.

An issue in many evaluations is that the quantitative data sets fall short of meeting assumptions underlying many of the available quantitative analysis techniques. The obtained data may not meet the assumptions of interval or ratio scales required by some statistical analysis techniques, and in many cases the employed data may have only marginal reliability and validity. Also, program participants rarely are selected randomly from a defined population. This complicates the aim of drawing inferences about some population of interest based on the obtained information. It also should be remembered that even when a random sample is successfully selected from a population of persons, that population very likely will be different by the time data have been gathered, analyzed, and reported. In many evaluations, the data respondents are the total population of interest, and there is no issue of using inferential statistics to generalize findings to a larger population of interest. Therefore, evaluators need to keep in mind and honestly report limitations and weaknesses in the data that underlie the quantitative analyses. They should employ inferential statistics only when they are relevant to the evaluation questions, and they should employ only those analysis techniques whose assumptions are at least minimally met by the data.

Quantitative Analysis Process

Evaluators should start the quantitative analysis process by exploring and gaining an understanding of the data set, identifying strengths and weaknesses in the data, making needed corrections, and discerning which of the desired questions can be addressed appropriately with the data. In this process, they should look for data that lie outside the bounds of reasonable expectation and appear to be in error. The point in identifying and analyzing such outliers is to confirm the validity of the information or disconfirm and delete or correct them.

Often a surfeit of data accumulates, which can too easily lead to fuzzy or even useless interpretations. Thus, the main aim of the quantitative analysis process is to reduce and

synthesize information so that the evaluation questions may be addressed rigorously and concisely. We deliberately have used the term *reasonable expectation* to give advance warning that data analysis must lead to interpretations that are credible to whoever proposed the questions that triggered the study. Members of this audience need to grasp the import of the gathered data and particularly how these data relate to their questions.

Evaluators should follow the start-up, exploratory analysis stage with more systematic, often increasingly complex analyses aimed at providing clear results and warranted interpretations. They should avoid, however, using complex statistical techniques when an audience would be served better by straightforward, simple methods. To help an audience understand and appreciate the results of analysis, evaluators should provide visual displays, such as cross-break tables, bar charts, and graphs, examples of which are readily available in appropriate texts; the evaluation magazine *Consumer Reports;* journal articles in the various disciplines; financial reports; and newspapers, such as *USA Today*. We add one further word of advice to evaluators who are uncertain about the relative importance and suitability of descriptive or inferential statistics: it is advisable to first explore the utility of descriptive statistics and graphics, such as the examples we have given. Whether inferential or statistical methods are used, it is essential that these are preceded by a thorough knowledge of collected data and their limitations.

Quantitative Analysis in Comparative Studies

Often evaluation audiences want to know whether one treatment is better than another or whether an innovative program is superior to an existing program. In such situations, evaluators may design an evaluation to compare different groups in different programs. For practical reasons, the comparison groups seldom are formed by random assignment, a problem that calls into question the equivalence of the groups.

Nonrandom assignment of subjects to comparison groups introduces a host of difficulties in discerning whether observed between-group differences in outcomes were due to differences in treatments. The different outcomes might reflect only original differences between the groups. Also, complications that impede interpretations of outcome differences arise when treatment and control groups are influenced differently, not only by the treatments they received but also by factors in their separate environments. As another example, differential dropout rates (often referred to as attrition) for experimental and control subjects might be as influential (or more so) in producing outcome differences as the administered treatment and control conditions. When there are no observed outcome differences, it is possible that the experimental treatment was not carried out as planned.

These examples of difficulties in conducting comparative studies underscore that quantitative analysis in such studies is a daunting task that requires care; resourcefulness; incisive investigation (and associated costs); multiple methods; documentation of procedures and difficulties; and a great deal of circumspection, caution, and candor in interpreting and reporting findings. And even when subjects are assigned to treatment and control groups randomly, many intervening factors as outlined earlier—such as differential attrition, inadequate implementation of treatment plans and control conditions, and contextual influences—can confound the

obtained outcome measures. In *Experimental and Quasi-Experimental Designs for Generalized Causal Inference,* Shadish, Cook, and Campbell (2002) enumerated threats to internal and external validity, as well as statistical conclusion and construct validity, in relation to a wide range of research and evaluation designs. To overcome such difficulties, evaluators often are wise to supplement their statistical analyses with descriptive, in-depth case studies of the experiences and outcomes of comparison groups.

As discussed in Chapter 11, threats to internal validity include the following (Shadish et al., 2002):

1. *Ambiguous temporal precedence*. Lack of clarity about which variable occurred first may yield confusion about which variable is the cause and which is the effect.
2. *Selection*. Systematic differences in respondent characteristics across conditions could cause the observed effect.
3. *History*. Events occurring concurrently with treatment could cause the observed effect.
4. *Maturation*. Naturally occurring changes over time could be confused with a treatment effect.
5. *Regression*. When units are selected for their extreme scores, they will often have less extreme scores on other variables, an occurrence that can be confused with a treatment effect.
6. *Attrition*. Loss of respondents to treatment or to measurement can produce artifactual effects if that loss is systematically correlated with conditions.
7. *Testing*. Exposure to a test can affect scores on subsequent exposures to that test, an occurrence that can be confused with a treatment effect.
8. *Instrumentation*. The nature of a measure may change over time or over conditions in a way that could be confused with a treatment effect.
9. *Additive and interactive threats*. The impact of a threat can be additive in relation to that of another threat or may depend on the level of another threat.

Equally relevant, but not discussed in Chapter 11, are Shadish et al.'s identified threats to external validity (2002), which are as follows:

1. *Interaction of the causal relationship with units*: An effect found when certain kinds of units might not hold if other types of units have been studied
2. *Interaction of the causal relationship over treatment variations*: An effect found when one treatment variation is not holding with other variations of the treatment, when that treatment is combined with other treatments, or when only part of a treatment is used
3. *Interaction of the causal relationship with outcomes*: An effect found on one kind of outcome observation but that might not hold if other outcome observations have been used
4. *Interaction of the causal relationship with settings*: An effect found in one kind of setting that may not hold in other settings
5. *Context-dependent mediation*: An effect found when an explanatory mediator of a causal relationship in one context is not necessarily mediating in another

Testing Statistical Hypotheses

For the majority of social science researchers, and for many evaluators, formulating hypotheses is the sine qua non of all social inquiry. In essence, a research hypothesis is a deductive guess that states an expected outcome of a study. When formulating a hypothesis, the evaluator deduces an anticipated result through a literature review process, experience, or observation. Hypotheses can be expressed in numerous ways, but typically a hypothesis is formulated first as a null or nil (literally meaning "zero difference" or "zero relationship") hypothesis; then as either an alternative nondirectional (two-tailed, two-sided) hypothesis or an alternative directional (one-tailed, one-sided) hypothesis. Alternative, nondirectional hypotheses imply that a difference is anticipated but do not express the direction of that difference. Directional hypotheses, however, state the expected direction of an expected difference. These types of hypotheses are presented and defined in Table 23.2. In general, many evaluators are interested in one of the alternative hypotheses, whether directional or nondirectional, not the null or nil hypothesis.

Null or nil hypotheses are present in nearly all forms of social science research and most evaluations, whether explicitly stated or not. And nearly all statistical tests are tests of null or nil hypotheses rather than tests of alternative hypotheses. Hypotheses can be expressed in numerous ways, and the methods for doing so vary based on disciplinary traditions, norms, and standards. Some biostatisticians, for example, refer to these types of tests of hypotheses as tests of equivalence (for example, Is a new drug that costs less to produce as effective as an older drug that costs more to produce?) or superiority (for example, Is a 500 milligram dose more effective than a 250 milligram dose of the new drug?).

If an evaluator is interested in determining whether a new reading curriculum is more effective or less effective than the existing curriculum (a nondirectional hypothesis), he or she might express the question, Is the new reading curriculum more effective than the existing curriculum? (a directional hypothesis claiming superiority) as a null hypothesis and as an alternative hypothesis. In notational form, where H_0 is the null hypothesis, H_A is the alternative hypothesis (where the subscript A represents the alternative), μ_T is the treatment mean (the new curriculum), and μ_C is the control mean (the existing curriculum), this hypothesis would be represented as follows:

$$H_0 : \mu_T = \mu_C$$

$$H_A : \mu_T \neq \mu_C$$

Table 23.2 Common Types of Hypotheses

Type of Hypothesis	Definition
Null or nil hypothesis	States that no difference is expected
Nondirectional hypothesis	States that a difference is expected, but does not state the direction of the expected difference
Directional hypothesis	States that a difference is expected, and states the direction of the expected difference

Using the same null hypothesis, the evaluator might formulate an alternative directional hypothesis, rather than a hypothesis simply suggesting that the two population means differ in an unknown direction, that specifies that the treatment mean, μ_T, is greater than the control mean, μ_C. This directional hypothesis would be expressed as follows:

$$H_A: \mu_T > \mu_C$$

Type I and Type II Errors

Two concepts are important considerations for understanding the practice of null hypothesis significance testing: Type I and Type II errors. A Type I error is the conditional prior probability of rejecting H_0 when it is true, where this probability is typically expressed as alpha (α). Alpha is a prior probability because it is specified before data are collected, and it is a conditional prior probability, *p,* because H_0 is assumed to be true. This conditional prior probability is usually expressed as

$$\alpha = p(\text{Reject } H_0 \mid H_0 \text{ true})$$

where — means "assuming" or "given." Both p and α are derived from the same sampling distribution and are interpreted as long-run, relative-frequency probabilities. Unlike α, however, p is not the conditional prior probability of a Type I error (often referred to as a false positive) because it is estimated for a particular sample result. The conventional level of α is either 0.05 or 0.01 in most of the social sciences (Cohen, 1994). Alpha is the risk of a Type I error, akin to a false positive because the evidence is incorrectly taken to support the hypothesis, for a single hypothesis only (sometimes referred to as a primary or focal outcome). When multiple statistical tests are conducted, there is also a family-wise (family-wise error [FWE]) probability of Type I error (sometimes referred to as multiplicity), which is the likelihood of making one or more Type I errors across a set of statistical tests. If each test is conducted at the same level of α, then

$$\alpha_{FWE} = 1 - (1 - \alpha)^C$$

where C is the number of tests performed, each at a specified α level. In this equation, the term $(1 - \alpha)$ is the probability of not making a Type I error for any individual test, is the probability of making no Type I errors across all tests, and the whole expression represents the probability of making at least one Type I error among all tests. So, for example, if ten statistical tests were performed, each at $\alpha = 0.05$, the family-wise Type I error rate would be

$$\alpha_{FWE} = 1 - (1 - \alpha)^{10} = 0.40$$

Thus, the Type I error rate across all ten statistical tests would be 40 percent. This result indicates the probability of committing one or more Type I errors, but it does not indicate how many errors have been committed or in which specific statistical test, or tests, the error occurred.

There are two basic ways to control family-wise Type I error: Either reduce the number of tests (or only test the primary or focal outcome) or lower α to a tolerable rate for each test.

The former strategy reduces the total number of tests to include only those with the greatest substantive meaning. Using the latter strategy, the α rate can be determined by a number of methods, including the Bonferroni correction. The Bonferroni correction simply requires dividing the target value of α_{FWE} by the number of tests, and setting the corrected level of statistical significance at α_B where

$$\alpha_B = \frac{\alpha_{FWE}}{C}$$

Therefore, if ten statistical tests were conducted and the tolerable Type I error rate was 5 percent, then for each individual test,

$$\alpha_B = \frac{0.05}{10} = 0.005$$

Although formal tests of statistical significance largely originated from the works of Fisher (1925) and Neyman and Pearson (1933), the concepts of statistical power and Type II error have been substantially advanced by Cohen (1969, 1980, 1988, 1994). Power is the conditional prior probability of making the correct decision to reject H_0 when it is actually false, where

$$\text{Power} = p(\text{Reject } H_0 \mid H_0 \text{ false})$$

A Type II error (often referred to as a false negative) occurs when the sample result leads to the failure to reject H_0 when it is actually false. The probability of a Type II error is usually represented by beta (β), and it is also a conditional prior probability:

$$\beta = p(\text{Fail to reject } H_0 \mid H_0 \text{ false})$$

because power and β are complimentary:

$$\text{Power} + \beta = 1.00$$

Therefore, whatever increases power decreases the probability of a Type II error and vice versa. Several factors affect statistical power, including the α level; sample size; score reliability; design elements (for example, within-subject designs, covariates); and the magnitude of an effect, among many others (Cohen, 1988; Lipsey & Hurley, 2009). By lowering α, statistical power is lost, thus reducing the likelihood of a Type I error, which simultaneously increases the probability of a Type II error. Conversely, increasing sample size generally increases power. The relationship between Type I and Type II errors arising from statistical hypothesis testing is summarized in Table 23.3.

Null or nil hypothesis significance testing, in the social sciences and in many other disciplines, has been widely misused and misinterpreted (for example, a p value is thought to be the probability that a result is due to sampling error, or a p value is thought to be the probability that a decision is wrong). The correct interpretation of p values, for $p < 0.05$,

Table 23.3 The Accept-Reject Dichotomy and Decisions for Hypotheses

Decision	H_0 Is True	H_0 Is False
Do not reject H_0	Correct decision $(1 - \alpha)$	Type II error (β)
Reject H_0	Type I error (α)	Correct decision $(1 - \beta)$

essentially includes only the following (Kline 2004, 2008; Thompson, 2006; Wilkinson and Taskforce on Statistical Inference, 1999):

- The odds are less than one in twenty of getting a result, from a random sample, even more extreme than the observed sample when H_0 is true.
- Less than 5 percent of test statistics are further away from the mean of the sampling distribution under H_0 than the one for the observed result.
- Assuming H_0 is true and the study is repeated many times, less than 5 percent of these results will be as inconsistent with H_0 as the observed result.

Effect Sizes and Practical Significance of Findings

Given that classical statistical significance testing provides only limited information—and is largely a function of sample size—determining practical significance, in part, requires estimating and reporting relevant effect sizes and confidence intervals. Effect sizes provide information about the direction and magnitude of an effect, and confidence intervals provide information about the precision of an estimated effect size (for detailed discussions of effect size interpretation, see Cooper, Hedges, and Valentine [2009]; Ellis [2010]; and Kline [2004]). Effect sizes can be computed in multiple ways, in unstandardized or standardized forms. Unstandardized effect sizes include raw mean differences (for example, Glass's Δ), which can be used on a meaningful outcome measure, such as blood pressure. When outcomes are measured on nonintuitive constructs (such as empathy), standardized means differences, such as Cohen's d or Hedges's g, can be applied. Other common effect size metrics include odds ratios and risk ratios for binary data (for example, live or die, pass or fail) and various measures of association (for example, r).

Using Cohen's conventions (1988), standardized effect sizes of 0.20, 0.50, and 0.80 are considered small, medium, and large, respectively. In the upper portion of Figure 23.3, the pretest distribution, or baseline (where μ_1 represents the population parameter of the mean) and posttest distribution (where μ_2 represents the population parameter of the mean) perfectly overlap and, therefore, Hedges's g is 0.00 (that is, there is no difference or no effect). In the middle and bottom parts of Figure 23.3, the pretest and posttest distributions are distinctly separate and represent a difference, or change, in the pretest and posttest distributions. In the middle part of Figure 23.3, the difference between the pretest and posttest distributions is equivalent to one-half of a standard deviation (that is, Hedges's g is 0.50).

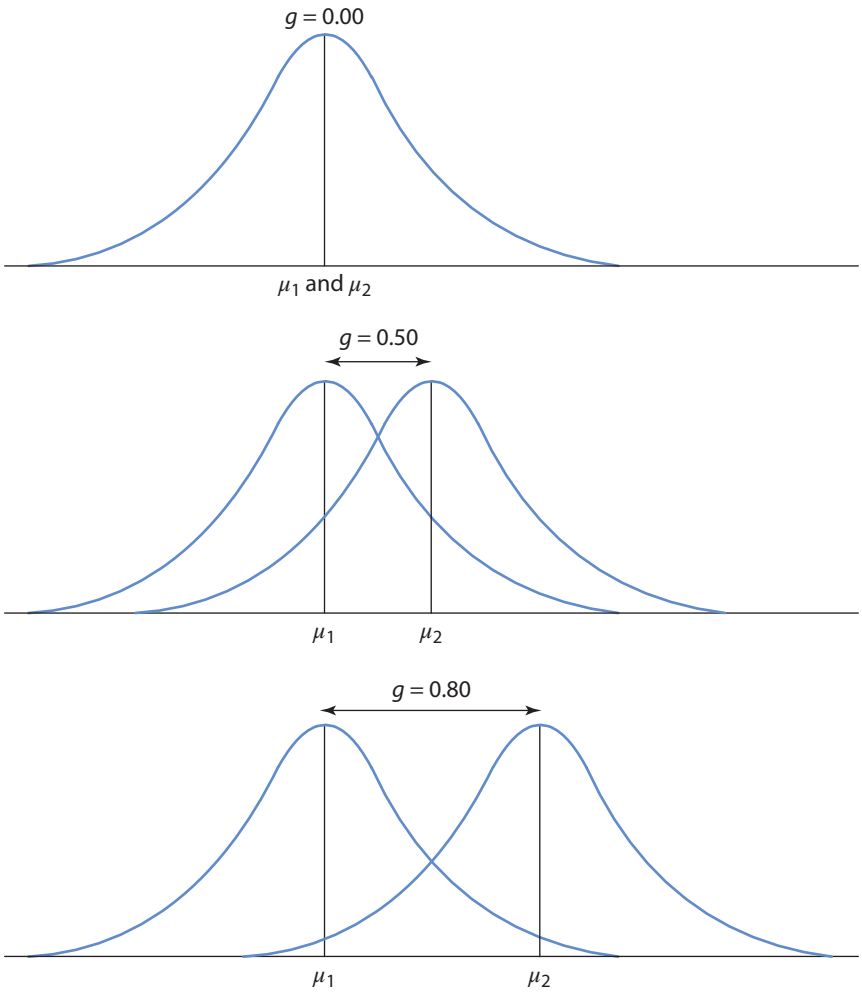

Figure 23.3 Hypothetical Examples of Hedges's *g* Effect Sizes

The primary advantage of reporting effect sizes in standard deviation units is that the results are metric-free, meaning they can be compared across outcomes. A secondary advantage of presenting effect sizes in this way is that variables that lack an intuitively meaningful metric (such as social capital or quality of life), unlike variables that are more widely understood (such as income or blood pressure), can be meaningfully interpreted.

Despite Cohen's conventions (1988), however, in the right context, even very small effects (for example, below 0.10) may be practically meaningful. This can happen in at least one of three ways. First, small effects can be meaningful if they are antecedents of larger consequences. Second, small effects can be important if they change the probability that larger effects will occur. Third, small effects can be important if they accumulate to become larger effects.

Equally important to reporting effect sizes is reporting confidence intervals (CIs), including the lower limit (LL) and upper limit (UL), usually in the form of a 95 percent CI. Essentially,

a CI provides information about the precision of an estimate, such as an effect size or other statistical estimator, or the degree of uncertainty associated with a sample estimate of a population parameter. A CI is therefore an interval estimate combined with a probability statement; that is, it is a representation of the confidence level, which is typically 90 percent or 95 percent. A point estimate, such as an odds ratio (OR) of 0.25, is a single value given as the estimate of a population parameter that is of interest. An interval estimate, such as a 95 percent CI, where the OR of 0.25 has an LL of 0.09 and a UL of 0.76, specifies instead a range that is estimated to include the parameter based on the obtained sample information, as shown in the random-effects model in Figure 23.4 (also see Chapter 6). Essentially, a confidence interval gives an estimated range of values that is likely to include an unknown population parameter, the estimated range having been calculated from a given set of sample data. For example, in the second study shown in the meta-analysis in Figure 23.4, the estimated OR is 0.63. In the study, the 95 percent CI has an LL of 0.34 and a UL of 1.16 (represented by the horizontal line running through the square). Given that the 95 percent CI crosses the null OR value of 1.00 (representing equal odds), the study is not considered statistically significant. Although the OR of 0.63 is practically significant, it is not considered statistically significant due to the imprecision of the estimate represented by the 95 percent CI. Conversely, the fifth study in the meta-analysis, having an OR of 0.18 (with an LL of 0.05 and a UL of 0.62), is equally imprecise (having a wide CI), but the CI does not cross the null value of 1.00, making the study statistically significant. CIs are preferred to point estimates because only CIs indicate the precision and uncertainty of an estimate. It is one thing to claim that the mean SAT score of

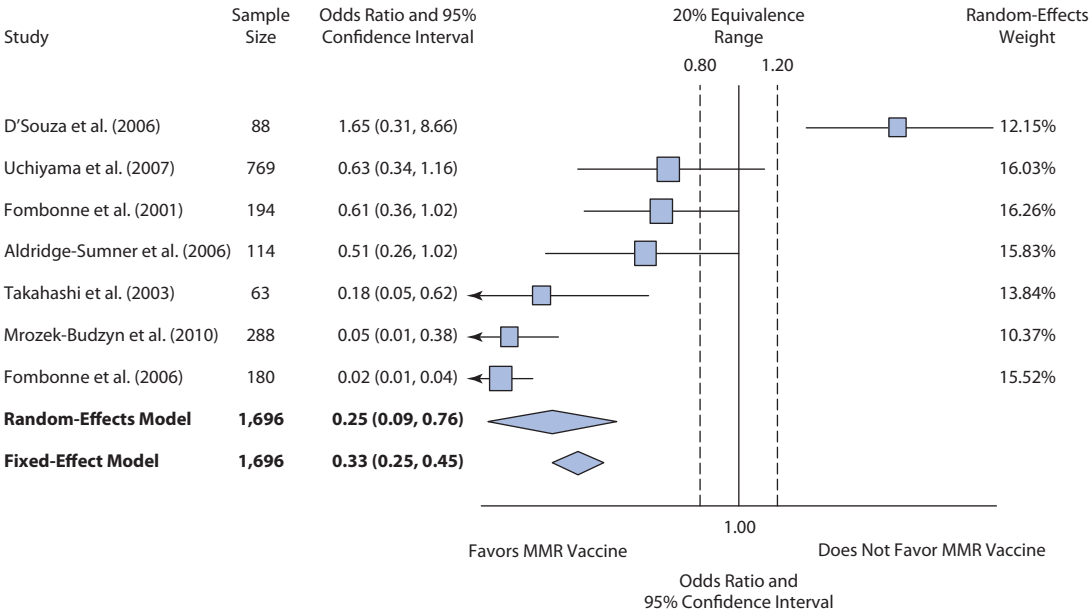

Figure 23.4 Meta-Analysis Forest Plot with a 20 Percent Equivalence Range

Source: Hobson, K. A., Mateu, P., Coryn, C.L.S., & Graves, C. D. (2012). Measles, mumps, and rubella vaccines and diagnoses of autism spectrum disorders among children: A meta-analysis. *World Medical & Health Policy*, *4*(3), 7.

students in a school is 600, but quite another to claim that the range of true scores is somewhere between 400 and 800 (or, stated differently, 600 ± 200).

Determining Consistent, Replicable Patterns of Results

As feasible, evaluators should employ different methods of analysis to determine whether a consistent and replicable pattern of results is present. They might apply parallel parametric and nonparametric techniques, quantitative as well as qualitative techniques, and interpretations against individual scores as well as group means. In many comparative evaluations, the treatment is applied to intact groups, and the group is the correct unit of analysis. However, given that the number of such treated groups often is small, the resulting analysis of group means will lack power for detecting small but important differences (and thus carry a high risk of Type II error). In such a case, the evaluator can gain statistical power by using individual scores to check for significant differences. In the area of exploring the data, we think such a check is appropriate, although the individual is not the statistically correct unit of analysis.

Use of multiple techniques in analyzing data can help an evaluator overcome some of the deficiencies in the data and help members of the evaluation's audience determine how much they should trust the reported findings. Evaluators should produce such overall statistics as group means, medians, and standard deviations, but also should look more deeply into the data. Determining only the average performance of program participants might mask important positive or negative effects on subgroups or individuals. In comparative analyses, evaluators often should follow up statistical tests of main effects with tests for statistical interactions and subsequent a posteriori tests for simple effects on subgroups. Such analytical methods as analysis of covariance might be used to adjust for initial differences between nonequivalent groups, but it is often difficult to meet the assumptions underlying these tests. Evaluators must not assume or imply to their audiences that such techniques as gain score analysis, matching, or analysis of covariance will necessarily adjust sufficiently for preexisting differences between comparison groups.

Documenting and Validating Quantitative Analyses

Ultimately evaluators should ensure that their quantitative analysis techniques and calculations are defensible. They should document the procedures they employed; state the assumptions required by these techniques; report the extent to which the assumptions were met; and justify (and, as appropriate, qualify) their interpretations of the results of their analyses. They should report potential weaknesses in the evaluation design or data analysis (for example, violation of scaling or randomization assumptions or program participants' dropping out) and discuss their possible influence on interpretations and conclusions. They should accord importance to both rigor and relevance and should not assume that statistically significant results are necessarily practically significant. Evaluators also should not credit nonsignificant statistical results just because stakeholders judge them as practically significant, and should only credit such results if there is evidence that the small but practically important difference is replicable. Again, we emphasize that evaluators and their clients should set the level of statistical significance in consideration of the potential for and importance of Type I and Type II errors.

Evaluators should bear in mind that quantitative analyses often fail to provide sufficient insight into the most important questions. A former president of a major university often assessed statistical comparisons of universities by saying, "Statistics are like bikinis; they reveal a great deal but always conceal the essentials." Although statistical analyses are often important, many require follow-up qualitative analysis.

As is evident in this section of the chapter, quantitative analysis is a highly developed technical discipline, including many esoteric terms, concepts, techniques, and formulas. In many evaluations the evaluator can meet the needs of the study by applying only rudimentary methods, such as computation of means, medians, percentages, standard deviations, correlations, *t*-tests, and analyses of variance. Other more sophisticated and complicated studies will require more advanced techniques, such as those listed earlier in this chapter. When such sophisticated techniques are required, the evaluator should consider engaging a specialist in statistics to conduct or guide the needed analyses and, more generally, to ensure their soundness. Often large-scale evaluations appropriately are team efforts, with a key member of the evaluation team being a highly competent statistician.

Analysis of Qualitative Information

In its 1994 edition of *The Program Evaluation Standards,* the Joint Committee stated, "Qualitative information in an evaluation should be appropriately and systematically analyzed so that evaluation questions are effectively answered" (p. 165).

Patton (1990) noted that a qualitative inquiry has a foundation built on several interconnected themes. In this section, we present some of these themes as we describe various strategies of qualitative inquiry:

- *Naturalistic inquiry*: A nonmanipulative study of situations as they unfold naturally.
- *Inductive analysis*: Immersion in details of data to delineate categories or sets of information and their interrelationships.
- *Holistic perspective*: Studying the whole phenomenon of an evaluand that may not be reduced to discrete variables (as occurs commonly with quantitative analysis).
- *Qualitative data*: Arising from and encompassing a range of techniques that capture perspectives and experiences through the evaluator's personal contact with study subjects and their actual situations.
- *Case study orientation*: Capturing the true nature of individual, unique cases. Endeavoring to be as objective as possible, the evaluator does not advance personal views or agendas. (For an in-depth description of case study methodology, see Chapter 12.)

It is difficult to imagine any evaluation study not including some qualitative information.

Evaluators typically acquire and analyze a wide range of qualitative information—for example, proposals and accountability reports; staff résumés; meeting minutes; correspondence files; beneficiaries' judgments of services; letters to the editor; site visit reports; participant observers' reports; case study reports; newspaper articles; public relations brochures; interview

responses; independent observers' field notes; oral testimony; written complaints; award documents; photographs; video- and audiotapes; focus group transcripts; public forum reports; proceedings of hearings; conference reports; various kinds of records; and unsolicited comments, accounts, and judgments. Qualitative information often is collected by design, but some of it may appear unexpectedly or be discovered through exploratory investigation. It may concern a wide range of program variables, such as beneficiaries' needs and wants, how and why a program got started, goals and plans, schedules and budgets, personnel and procedures, equipment and facilities, operations and expenditures, and intended and unintended outcomes. Descriptive studies—involving, for example, documentation of a program's activities, definition and description of the program's stakeholders, and description of staff credentials—rely heavily on qualitative information.

When qualitative information is collected, evaluators should employ systematic, rigorous methods of qualitative analysis. The Joint Committee (1994) defined qualitative analysis as "the process of compiling, analyzing, and interpreting qualitative information about a program that will answer particular questions about that program" (p. 171). Qualitative analysis culminates in narrative presentations, such as a summary of main outcomes, a discussion of the extent to which program plans were well executed, a depiction of major and minor themes running through stakeholder inputs, identification of inconsistencies as well as consistencies in different sets of obtained information, a contrasting of findings from different stakeholder viewpoints, a contrasting of findings at different points in time, and interpretations of cause-and-effect relationships.

Gathering qualitative information has many benefits in an evaluation. These benefits include providing breadth of perspective and depth of information, buttressing and complementing quantitative findings, confirming or disconfirming quantitative findings, rounding out the full countenance of a program, and helping the audience perceive a program's essence and nuances. Also, pertinent quotations may be reported along with quantitative results. In qualitative analyses, it is essential to consider alternative and possibly conflicting perceptions of reality as well as different values from which to judge programs.

Qualitative Analysis as a Discovery Process

In contrast to quantitative analysis, qualitative analysis often evolves in a process of discovery rather than following a predetermined analysis plan. In the course of qualitative analysis, evaluators often have to generate information collection devices, category systems, and methods of summarizing and displaying information throughout an evaluation. Whereas quantitative analysis typically focuses on information that was collected from a predetermined sample, qualitative analysis often uses information from snowball samples that grow and take direction based on successive exchanges with key informants. As the Joint Committee (1994) stated,

> Qualitative analysis often involves an inductive, interactive, and iterative process whereby the evaluator returns to relevant audiences and data sources to confirm and/or expand the purposes of the evaluation and test conclusions. It often requires an intuitive sifting of expressed concerns and relevant observations. (p. 171)

In applying qualitative analysis techniques, evaluators should allow emergent questions to shape the collection and analysis of qualitative information as an evaluation proceeds. For each set of qualitative information, evaluators should choose an analytical procedure and plan for summarizing findings that are appropriate for addressing part or all of the evaluation's questions and that suit the nature of the information to be analyzed. They should define the boundaries of the information to be examined in terms of such components as targeted beneficiaries, geographical location, time period to be examined, financial sponsors, and program budget. By identifying pervasive themes in the information, evaluators should ferret out meaningful categories of information, such as innovative methods, undue control by administrators, democratic leadership, motivated staff, personality conflicts, value conflicts, inadequate (or adequate) supervision, goal drift, and community involvement. In communicating findings to audiences, evaluators might extract certain findings from the qualitative analysis and embed them in the presentation of quantitative findings.

Practical Steps in the Qualitative Analysis Process

Initially it can be useful to analyze separately the information obtained from each qualitative method—for example, interviews, open-ended questionnaires, focus groups, or documents. Each such set of information might be examined to address evaluation questions concerning such matters as beneficiaries' needs; program implementation; intended effects; side effects; and judgments of quality, utility, and significance.

Criteria for Judging Qualitative Analyses

In general, any one set of qualitative information has been sufficiently and appropriately analyzed when the following are true:

- The evaluator has derived a set of categories that unambiguously account for the obtained information and amplify and address the evaluation questions.
- The information has been parsimoniously grouped into categories.
- The categories of information have been verified as reliable and valid.
- The categories have been applied to produce meaningful inferences and conclusions.
- The qualitative analysis process has been documented and validated.
- The evaluator has forthrightly reported any potential weaknesses in the information and its analysis.

General Process for Analyzing Qualitative Information

The general process we have found useful in analyzing given sets of qualitative information can be summarized in the following steps:

1. Compile a set of documents for each type of qualitative information, such as correspondence, newspaper clippings, transcripts of focus group meetings, and notes from interviews.

2. Mark each document in each set with a unique identification number (for example, for correspondence, Cor-1, Cor-2, Cor-3, and so on).
3. Read through a random sample of the documents in each set, making marginal notes on points that seem relevant to the evaluation's purposes and questions. Relevant information might pertain, for example, to characteristics of beneficiaries, needs of beneficiaries, how and why the program was launched, strengths of the program design, innovative methods, program detractors, staff competence, indications of graft, program implementation, program costs, program outcomes, and side effects. Such marginal notes provide a grounded basis for generating categories for use in the qualitative analysis. Note that these are not preconceived categories but rather groupings that come to mind when first studying random samples of materials in each set.
4. In each set of materials, group the marginal notes into an efficient set of categories to eliminate minor, trivial differences between categories across voluminous marginal notes.
5. Contrast the derived sets of categories, and synthesize them into a coherent set of categories that is faithful to what was obtained for each set of qualitative information and is as efficient as possible.
6. Contrast the derived set of categories with the conceptual framework and main questions guiding the evaluation, and develop a standardized set of categories for the subsequent analysis of qualitative information. This finalized set of categories should reflect the previous empirically derived categories, the guiding evaluation approach, and the evaluation's main questions. This is the stage in which an evaluation approach such as Scriven's Key Evaluation Checklist (see Chapter 14); Stake's countenance approach (see Chapter 15); or Stufflebeam's context, input, process, and product (CIPP) model (see Chapter 13) can be especially useful.
7. Apply the standardized set of categories to analyzing each set of qualitative information. Continue reading the material in each set until a relevant category has been attached to each noteworthy segment of each document.
8. For each set of information, summarize what has been learned in relation to each category of findings—for example, the program's costs, community support and opposition, conflicts of interest, main effects, and side effects. Also annotate the summary with the identification numbers of the relevant source documents. This is important preparation for answering later questions from recipients of the final evaluation report.
9. Looking across the summaries for the different sets of information, write findings in relation to each evaluation question—for example, To what extent did the program reach all the intended beneficiaries? To what extent were the outcomes worth the effort and cost? To what extent was the program institutionalized? This procedure will help clarify issues related to evaluation questions, as well as help shape the content and nature of the final report.
10. Subject the results of the qualitative analysis to independent critiques, and resolve any identified deficiencies.

Information on qualitative analysis techniques, including uses of computer software, is available in a wide range of publications. Some examples are Crabtree and Miller (1992); Denzin and Lincoln (2005); Fetterman (1998); Fielding and Lee (1991, 1998); LeCompte and Goetz (1982); Leninger (1985); Mabry (2003); Miles and Huberman (1984); Patton (1987, 1990); Strauss (1987); Tesch (1990); Wolcott (1995); and Yin (1991).

Qualifications Needed to Conduct Qualitative Analyses

Those who practice qualitative analysis need appropriate training and an appreciation for rigor as well as relevance. They should be proficient in such tasks as interviewing stakeholders, developing focus group questions, recording fieldwork data, interpreting historical information, taking accurate notes, conceptual analysis, text analysis, computer-assisted content analysis, historical analysis, videotape analysis, audiotape analysis, coding and classifying information, grounded theory analysis, and writing qualitative research reports (that take into account any applicable outcomes from quantitative research). Especially, they should be adept at identifying themes and majority and minority positions in a body of information.

Errors to Avoid in Analyzing Qualitative Information

Despite the importance of qualitative analysis, evaluators should not become overzealous in conducting qualitative analyses. They must not get carried away with the emergent, divergent, and in-depth features of qualitative analysis. They should not overstress the details of program circumstances and as a consequence obscure more general, pervasive findings that are likely to be of interest to their audience. They should not be so enticed by interesting new questions that they neglect to address the evaluation's main questions. They should be judicious and parsimonious in collecting qualitative information so that they do not make the evaluation too expensive and time consuming. Moreover, we cannot overemphasize that quantitative information and qualitative information are complementary and should work together to support the evaluation's findings and conclusions.

Validating Qualitative Analyses

Whatever methods of qualitative analysis are employed, evaluators should ensure the accuracy of findings by seeking confirmation from quantitative information, and they should verify the resulting inferences and conclusions. They should judiciously examine different sources of evidence on such bases as verifiability, credibility, and the degree of evaluator contact with the assessed entity. Evaluators should closely examine the validity of preconceptions, working hypotheses, generally accepted past practices and beliefs, and cited past evaluative conclusions. To test the consistency of categories, themes, and conclusions, it is a good idea, whenever possible, to engage two or more independent evaluators to analyze the same set of information. Also, it is good practice to subject qualitative analysis results to an independent audit. In addition, evaluators should engage representatives of stakeholders to review and assess the validity and meaningfulness of drafts of qualitative analyses. Ultimately evaluators

should document the qualitative analysis process and report this documentation along with the results of having the analysis validated.

In general, the purpose of qualitative analysis is to enrich an evaluation's message and prevent invalid conclusions. When evaluators meet the Joint Committee's Qualitative Analysis standard (1994), they avoid using inappropriate methods of analysis; carefully document, cross-check, and evaluate their findings; prevent their audience from reaching premature closure or misinterpreting the results; and keep the evaluation within reasonable bounds of time and cost.

Justified Conclusions and Decisions

The Joint Committee (2011) stated, "Evaluation conclusions and decisions should be explicitly justified in the cultures and contexts where they have consequences" (p. 165).

Quantitative and qualitative analyses are intended to provide bases for reaching justified conclusions and decisions, which are the evaluation's final judgments and recommendations. The evaluator's bottom-line conclusions offer audiences a foundation for judging and making decisions about a program or other object of interest. The evaluation's conclusions must be carefully derived and shown to be sound. They must be both defensible and defended. They should be appropriately qualified in terms of the applicable time periods, contexts, activities, persons, purposes, and supporting evidence.

Evaluators should base their conclusions on all pertinent information collected; on appropriate analyses and logic; and on a systematic, defensible synthesizing process. They should show how this information relates to the conclusions. In reaching conclusions about a program's effectiveness, evaluators should identify side effects as well as main effects. As feasible, evaluators should present not only their bottom-line conclusions but also plausible alternative conclusions along with an explanation of why they were rejected. On the one hand, although evaluators should attempt to address the audience's questions, they should be careful not to present conclusions that extend beyond the limits of their data. On the other hand, they should not be overly cautious in interpreting the evaluation's findings. A report that leads to effective decision making is devoid of exaggeration and pretension, but replete with justified statements of the evaluand's merit and worth.

In justifying conclusions, evaluators should supply the audience with full information about the evaluation's design, procedures, information, analyses, synthesis, and underlying assumptions. As feasible, evaluators should solicit feedback from a range of program stakeholders concerning the clarity and credibility of conclusions and recommendations. As appropriate, they should advise their audiences of any equivocal findings in the evaluation report and warn them to be cautious in applying those findings. Faulty and unexplained conclusions or ones that reach beyond the data may mislead audiences or cause them to disregard the evaluation.

The Synthesis Process

A key process related to the Joint Committee's Justified Conclusions and Decisions standard (2011) is that of synthesis: combining the study's value base, information, and analyses into a unified set of conclusions. In line with the mixed-method approaches advocated in this book,

and particularly in this chapter, the synthesis process involves information arising from both quantitative and qualitative inquiries.

The synthesis component is a highly challenging activity. It requires a determination of whether the audience requires a final synthesis; critical review of the available information and analyses to determine whether a final synthesis is feasible; rigorous application of logic and justifiable decision rules in relating the findings to evaluation questions and bottom-line areas of judgment; creativity in conceptualizing pertinent judgments; pragmatic thought plus reference to supporting evidence in developing actionable recommendations (if such are warranted); solicitation and use of critical reactions to draft judgments and recommendations; and proficiency in writing clear, substantiated, and properly qualified judgments and recommendations. Also, the evaluator should support the synthesis of evaluation findings with a detailed technical appendix or technical report that documents the evaluation design, information, and quantitative and qualitative analyses (see Chapter 24).

In the synthesis process, evaluators should focus primarily on the audience's questions and issues concerning the program's value. They should draw together relevant quantitative and qualitative analysis results pertaining to each evaluation question and areas for judgments and recommendations. The CIPP model provides a convenient advance organizer for grouping the audience's questions and the essential elements of a sound, comprehensive set of evaluative conclusions. The model's generic questions pertain especially to beneficiaries' needs, appropriateness of the program's plan and budget, adequacy of program implementation, reach to the intended beneficiaries, the amount and quality of outcomes, side effects, sustainability of the program, and transportability of the program. Main categories of bottom-line judgments are merit, worth, significance, and probity. In the context of the CIPP model, a good synthesis will provide an informative, justified response to each of these matters. With each such response, the evaluator might start with the quantitative or qualitative results and subsequently buttress these results with the other type of information. The write-up of each conclusion, as appropriate, should include areas of agreement across information sources, but it should also point out areas of contradictory evidence. Moreover, the synthesis process should be documented and subjected to independent assessment.

Scriven (1994b) noted that a final synthesis is not always needed or feasible (see Chapter 14). He also stressed that evaluators must not recklessly state a conclusion based only on personal judgment rather than on a logical link to solid, relevant evidence. Nevertheless, he stated that an evaluator should proceed toward a final synthesis if the client requires one, going only so far in that direction as is technically defensible.

Steps for Synthesizing Quantitative and Qualitative Information

We suggest that evaluators, to synthesize obtained evaluative information and reach defensible conclusions, carry out the following steps, which are roughly but not totally consistent with the steps recommended by Scriven (1994b, 2007):

1. Compile evidence on the assessed needs of the program's targeted beneficiaries and assess whether program goals are reflective of the assessed needs. If the answer is affirmative, the goals can be used as criteria for assessing the worth of outcomes. If the answer is negative, then the assessed needs should be employed to assess the worth of program outcomes.

2. Determine appropriate rules for reaching justified conclusions. A few examples are that a housing program for the working poor is at least partially meritorious if the different aspects of house construction passed all official city inspections; that the program is at least partially worthy if the projected number of members of the targeted group of working poor obtained high-quality houses and if at least 90 percent of them lived in the housing and kept up mortgage payments for at least four years; that the program has significance beyond the local application if it was replicated by other community development groups; and that it meets specified probity requirements if its books were audited and there were no indications of fraud or graft, and if it got a good report from program supervisors on its ethical treatment of program participants.
3. Select or derive defensible criteria for applying the decision rules. Example criteria of merit include the codes city inspectors use to approve electrical and plumbing installations; criteria of worth include those used to determine housing needs of the targeted beneficiaries plus the program's goals if they reflect assessed needs of the targeted beneficiaries; criteria of significance include the facts of successful replications of the project; and criteria of probity include the professional standards of the auditing and accounting fields.
4. Retrieve appropriate quantitative and qualitative evidence for applying the determined criteria of merit, worth, significance, and probity. Evidence of merit in the housing example could include city inspectors' reports and approval or disapproval of different aspects of each constructed house. Evidence of worth could be records of the program's beneficiaries' residing in their house over time, nurturing their children, caring for their property, meeting mortgage payments, and contributing to the health of their community. Evidence of program significance could be reports of site visits to projects that successfully replicated the subject housing project. Note that the evaluator would use results of both quantitative and qualitative analyses together in applying the evaluative criteria.
5. Determine if there is reliable and valid evidence for a sufficient range of criteria of merit, worth, significance, and probity to proceed with a determination of justified conclusions. Such a determination may be made for each of the four dimensions of value. It is possible that the client group might not be interested in the dimension of significance, and although no particular data may have been collected on probity, there may be no issue in this area. In most cases, however, a decision to proceed with the synthesis task should be supported by adequate evidence at least in the areas of merit and worth. At this point, the effort to synthesize information and reach justified conclusions appropriately should be aborted if there is not adequate evidence to proceed. This could be the case when the client and evaluator have previously agreed that the evaluation would be a limited effort focused, for example, on only a few formative evaluation issues.
6. Determine with stakeholders if the criteria of merit, worth, significance, and probity should be weighted differently. For example, some criteria in each set might be weighted as essential, whereas others could be weighted as important. In the ensuing analysis, the evaluator should judge the program a failure if it failed to meet any criterion designated as essential, no matter how well the program performed against the other criteria.

7. Use relevant evidence to rate the program, for example as 4 (strong), 3 (adequate), 2 (weak), 1 (unacceptable), or 0 (unratable), on each criterion. It is desirable to engage multiple raters who have successfully completed training and calibration and who have thoroughly studied the relevant evidence to accomplish this rating task.
8. Develop a bar graph of the rating results for each involved dimension (for example, merit, worth, significance, and probity). Each bar graph should array the employed criteria, identify those that are essential, and provide a bar reflecting the score for each criterion.
9. Provide a narrative conclusion for each dimension of value about the extent to which the evaluand satisfied the associated criteria and justify the conclusion with reference to the supporting quantitative and qualitative evidence. An evaluator should judge a program as failing on any dimension for which an essential criterion was not met.
10. Write an overall summary statement, considering all the dimensions of value, that assesses the evaluand's value. This statement should be prepared essentially in the form of an executive summary that reviews the synthesis process, presents the main conclusions, documents the decision rules, references the supporting evidence, and references the obtained independent assessments of the conclusions.
11. Determine whether useful recommendations can be presented and justified with supporting evidence. Recommendations should not be tendered if they are only intuitive. In some cases, the obtained evidence and relevant literature can be used to offer relevant, defensible recommendations. In other cases, it may be appropriate to propose a follow-up study designed explicitly to identify and evaluate alternative courses of action; related to the CIPP model, such a recommendation is tantamount to proposing that the client contract for a follow-up input evaluation.
12. Ensure that the final report's technical appendix or a separate technical report includes documentation—of such matters as the evaluation's questions, personnel and their qualifications, instruments, data collection and analysis procedures, synthesis steps, and verification provisions—to support the evaluation's conclusions and, if provided, recommendations.

Special Synthesis Procedures

Davidson (2005, 2011) has advocated using rubrics to derive nonarbitrary and defensible bottom-line evaluative conclusions, stressing that such rubrics can accommodate both quantitative and qualitative information simultaneously.

Gugiu (2007, 2011) has developed a more sophisticated method, which he has labeled "summative confidence," that relies on statistical and measurement theories in synthesizing information into evaluative judgments. Summative confidence is a statistical method for estimating a summative evaluative conclusion by accommodating the following: the Type I error rate; the number of, variance across, and correlation among the criteria used to formulate the conclusion; the performance benchmarks for critically important criteria; the sample size and the amount of measurement error for each criterion; and the amount of

weight accorded to each criterion. Gugiu's mathematically complex method also allows for determining the precision of a summative value judgment by placing a confidence interval around an evaluative conclusion.

Bottom-Line Steps in Producing Justified Conclusions

Essentially, justified conclusions are to be arrived at through a sequence of steps, such as the following:

1. Address each contracted evaluation question based on information that is sufficiently broad, deep, reliable, contextually relevant, culturally sensitive, and valid.
2. Derive defensible conclusions that respond to the evaluation's stated purposes (for example, to identify and assess the program's strengths and weaknesses, main effects and side effects, cost-effectiveness, and merit and worth).
3. Limit conclusions to the applicable time periods, contexts, purposes, and activities.
4. Identify the persons who determined the evaluation's conclusions (for example, the evaluator using the obtained information, plus the broad range of stakeholders who provided inputs).
5. Identify and report all important assumptions, the interpretive frameworks and values employed to derive the conclusions, and any appropriate caveats.
6. Report plausible alternative explanations of the findings and explain why rival explanations were rejected.
7. Document the entire process and its particulars for independent review.

Summary

Evaluation relies on principles of research—such as quantitative analysis and qualitative analysis—but also requires analysis of values. Because the state of the art in the latter area is primitive compared to that in quantitative and qualitative research methodology, some have advised evaluators to collect, analyze, and report only solid evidence. Accordingly, they have recommended that evaluators leave matters of synthesis and interpretation of findings to an evaluation's client. Such conservative, technically oriented practice is intended to keep evaluators from advancing into areas where they might make mistakes, but it falls short of being evaluation—the assessment of value. In this chapter we have departed sharply from the value-free line of advice. We believe that evaluators should make the client and other stakeholders parties to the synthesis process, especially in clarifying decision rules, criteria, and weights for criteria. Moreover, we posit that the essence of a professional evaluator's role is to conduct quantitative analysis, qualitative analysis, and synthesis toward the goal of reaching bottom-line, values- and evidence-based judgments and, as warranted, providing actionable recommendations.

This chapter has provided a general orientation to—and detailed explanations and advice for—conducting quantitative and qualitative analyses and reaching justified evaluative conclusions. We included definitions and examples of descriptive, relational, and causal questions and clarified the distinction between probabilistic and deterministic reasoning. Other parts of

the chapter's discussion of quantitative analysis touched on techniques and software packages for analyzing data, threats to internal and external validity (especially in comparative studies), hypothesis testing, Type I and Type II errors, confidence intervals, and calculating effect sizes to assess a program's practical significance. We subsequently discussed qualitative analysis and its complementary relationship to quantitative analysis. Whereas quantitative analysis was shown to be largely preplanned, qualitative analysis was depicted as an evolving, interactive process. We identified types of qualitative information to be analyzed, defined steps for carrying out qualitative analyses, defined criteria for judging such analyses, identified errors to avoid, and listed qualifications for conducting qualitative analyses. The chapter's final section focused on evaluation's culminating process of synthesizing the results of qualitative and quantitative analyses and its end goal of producing justified conclusions. This concluding section defined a process for synthesizing findings, discussed the nature of different types of cut scores, and listed bottom-line steps for producing justified conclusions.

REVIEW QUESTIONS

1. Construct a matrix, and fill in the cells to define, compare, and contrast the concepts of quantitative analysis, qualitative analysis, and synthesis of evaluation findings.
2. List the practical steps you would follow in analyzing a set of quantitative information, and explain why it is important to begin this process by exploring the data.
3. Identify some of the reasons why it is important for evaluators to employ multiple analysis techniques, including problems an evaluator could encounter while obtaining information most relevant to the evaluation questions. What are the benefits of employing multiple analysis techniques?
4. Define the concepts of Type I and Type II errors, explain why these concepts are important in evaluation work, and give three instances in which a Type II error would be more important than a Type I error.
5. In comparative analyses, why should evaluators often follow up statistical tests of main effects with tests for statistical interactions and subsequent a posteriori tests?
6. Compare and contrast the concepts of statistical significance and practical significance, give an illustration of why a finding might be statistically significant but lack practical significance, and identify some ways that an evaluator can examine a finding's practical significance. Then state whether it is possible for a finding to be statistically nonsignificant but practically significant, and justify your response.
7. List criteria for judging whether a set of qualitative information has been sufficiently and appropriately analyzed.
8. What is meant by the term *evaluative conclusions*? What is their role in an evaluation? How are evaluative conclusions appropriately justified, and what kind of information should an evaluator present to a client to justify a set of evaluative conclusions?

9. Characterize the process involved in synthesizing sets of quantitative and qualitative information. List as many steps as you can from this chapter's recommended process for synthesizing obtained evaluative information.
10. What is the rationale for Scriven's position that a final synthesis is not always needed or feasible? How does this claim relate to his concepts of formative evaluation and summative evaluation (see Chapter 14)?

Group Exercises

Exercise 1

Critique the following statement: "All evaluations should include convergent as well as divergent stages, and every evaluation should culminate in a bottom-line set of conclusions."

Exercise 2

Discuss and defend the claim that quantitative analysis and qualitative analysis are complementary processes. In your discussion, explain and illustrate how these processes support each other in the presentation of findings.

Exercise 3

Identify a comparative study that is familiar to at least one member of your group. Discuss this study in terms of what alpha and beta levels make sense given the possibility of Type I and Type II errors. Identify examples of Type I and Type II errors where the Type II error is more important than the Type I error. What are the implications of these examples for setting the alpha level for Type I error?

Exercise 4

Suppose your group has obtained a set of interview responses to a standard set of questions from different interviewees. List the steps your group would follow to analyze this set of qualitative information, with special reference to possible pitfalls you might encounter and must consider.

Exercise 5

Suppose your group has been assigned the task of developing a short course to train evaluators in the procedures of qualitative analysis. What procedures might you include in this course? Place the identified procedures into three groups: essential, important, and marginally important. (Group members could address this question individually, and then compare and discuss answers.)

Exercise 6

Discuss the following questions:

1. Under what conditions can an evaluator appropriately present a client with a set of recommendations?
2. Under what conditions is this not appropriate?
3. If the evaluator's basis for offering recommendations is too weak but the client still wants them, what course of action available from the CIPP model might the evaluator suggest to the client?

Suggested Supplemental Readings

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Skokie, IL: Rand McNally.

Crabtree, B. F., & Miller, W. L. (1992). *Doing qualitative research*. Thousand Oaks, CA: Sage.

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2005). *The Sage handbook of qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.

Fetterman, D. M. (1998). *Ethnography: Step by step* (2nd ed.). Thousand Oaks, CA: Sage.

Fielding, N. G., & Lee, R. M. (1991). *Using computers in qualitative research*. Thousand Oaks, CA: Sage.

Fielding, N. G., & Lee, R. M. (1998). *Computer analysis and qualitative research*. Thousand Oaks, CA: Sage.

Goldstein, H. (1987). *Multilevel models in educational and social research*. New York, NY: Oxford University Press.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics* (5th ed.). Boston, MA: Houghton Mifflin.

Hopkins, K. D., & Glass, G. V. (1978). *Basic statistics for the behavioral sciences*. Upper Saddle River, NJ: Prentice Hall.

Jaeger, R. M. (1990). *Statistics: A spectator sport* (2nd ed.). Thousand Oaks, CA: Sage.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Joint Committee on Standards for Educational Evaluation. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

LeCompte, M. D., & Goetz, J. P. (1982). Problems of reliability and validity in ethnographic research. *Review of Educational Research, 52*, 31–60.

Leninger, M. (Ed.). (1985). *Qualitative research methods in nursing*. Orlando, FL: Grune & Stratton.

Mabry, L. (2003). In living color: Qualitative methods in educational evaluation. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 167–188). Norwell, MA: Kluwer.

Miles, M. B., & Huberman, A. M. (1984). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (1987). How to use qualitative methods in evaluation. In J. L. Herman (Ed.), *Program evaluation kit* (2nd ed.; Vol. 4). Thousand Oaks, CA: Sage.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Thousand Oaks, CA: Sage.

Strauss, A. (1987). *Qualitative analysis for social scientists*. Cambridge, UK: Cambridge University Press.

Tesch, R. (1990). *Qualitative research: Analysis types and software tools*. Bristol, PA: Falmer.

Wiersma, W., & Jurs, S. G. (2005). *Research methods in education: An introduction* (8th ed.). Needham Heights, MA: Allyn & Bacon.

Winer, B. J. (1962). *Statistical principles in experimental design*. New York, NY: McGraw-Hill.

Wolcott, H. F. (1995). *The art of fieldwork*. Walnut Creek, CA: AltaMira.

Yin, R. K. (1991). *Case study research: Design and methods*. Thousand Oaks, CA: Sage.

CHAPTER 24

COMMUNICATING EVALUATION FINDINGS

Previous chapters have presented theories, standards, approaches, and procedures needed to conduct sound, effective evaluations. An effective evaluation will inform intended users about a program's merit and worth and stimulate them to make warranted decisions and needed improvements. It will also assist them in meeting needs related to program accountability. Despite the thoroughness of planning and executing sound data collection and analysis procedures, however, an evaluator may fail to effectively communicate findings to the full range of intended users. Unless interim and final findings are presented to an evaluation's intended users in a timely, systematic, convincing, ethical, and easily understood manner, the consequences of an excellent data-gathering and analysis process are likely to be nil or even counterproductive—especially due to wasted time and resources and discrediting of the evaluation function. Evaluators and their clients therefore must devote serious attention to the reporting function and excel in its execution.

The importance of effectively communicating evaluation findings to the full range of intended users cannot be overemphasized. Attention to reporting is an integral part of planning, even before data are gathered. During this early stage, the format, style, and content of the final report should be worked out using inputs from the full range of intended users of evaluation findings. Evaluators must be sensitive and responsive, especially to the program staff's information needs. Written and oral reports to staff should be clear, timely, and useful for program improvement. As appropriate, some of the evaluation reports should address particular questions of interest to intended users who are

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **Complex needs and challenges associated with reporting evaluation findings**
- **Procedures for determining and prioritizing an evaluation's intended users and identifying their intended uses of findings**
- **The importance of addressing in contractual agreements reporting and dissemination of findings, plus follow-up assistance for applying findings**
- **The need to engage and build trust among the full range of evaluation stakeholders, and ways to do so**
- **The role and composition of evaluation review panels**
- **Keys to effectively providing interim feedback**
- **Formats for final evaluation reports**
- **The contents of and procedures for applying the Evaluation Report Layout Checklist**
- **The applicability of visual processing theory in reporting**
- **Ways to lead discussions and manage conflict**
- **Suggestions for helping policymakers and administrators apply findings**
- **Issues and suggestions in regard to the presentation of recommendations**
- **The role of input evaluation in defining and assessing alternative courses of action based on an evaluation**

not part of the program's staff. By addressing the information needs of the full range of intended users of evaluation findings, chances are enhanced that the reports will be effectively received and used.

This chapter provides an orientation to and practical information for equitably and skillfully involving stakeholders in the evaluation process, addressing psychological and political threats to an evaluation's success, and effectively reporting and promoting the use of sound evaluative information. Our presentation is designed to be consistent with recent editions of the Joint Committee on Standards for Educational Evaluation's *Program Evaluation Standards* (1994, 2011), two of the most comprehensive and authoritative sources of practical advice for ensuring that evaluation findings are used. These standards documents stress and explain that evaluation reports should be relevant, sufficiently comprehensive, balanced, clear, timely, impartial, defensible, politically viable, and effectively delivered and disseminated. The Evaluation Impact standard states that "evaluations should be planned, conducted, and reported in ways that encourage follow-through by stakeholders, so that the likelihood that the evaluation will be used is increased" (Joint Committee, 1994, p. 59).

We have divided this chapter into six major sections: (1) points from previous chapters that are relevant to communicating evaluation findings, (2) complex needs and challenges in reporting findings, (3) ensuring conditions to foster use of findings, (4) providing interim evaluative feedback, (5) preparing and delivering a final report, and (6) providing follow-up assistance to enhance an evaluation's impact. Within each section, we summarize and elaborate relevant information from selected Joint Committee (1994, 2011) standards, add insights and examples gleaned from our evaluations, and use the discussion to generate practical advice. Sample interim and final evaluation reports produced by us and by our colleagues are available from www.josseybass.com/go/evalmodels.

Review of Pertinent Analysis and Advice from Previous Chapters

Basic aims of a good evaluation as presented in Chapters 11 through 16 are to stimulate and even excite stakeholders such that they gain insight into their own program, and to enlighten all involved in the program about its merit, worth, and areas for improvement. A program evaluation should inform, educate, and convince decision makers about various pathways and choices for strengthening their program. An evaluator should describe and assess a program in such a way that program leaders will be helped to decide whether to sustain, expand, reduce, or terminate the program. Sound evaluations that are effectively reported may assist with making decisions about purchasing and adopting a new program or adapting a current one. Moreover, evaluators should configure their reports to help program administrators address accountability needs, especially when the assessed program involves large outlays of funds. Finally, field-based program evaluations may inform a wider public about a program's background, structure, cost, implementation, and outcomes. In such applications, evaluation is a public service. Communication and reporting skills are essential to meeting the full range of evaluation purposes and uses. Evaluators must possess and apply such skills to effectively deliver findings during and following program implementation to aid program decision making and accountability.

Complex Needs and Challenges in Reporting Evaluation Findings

Inadequate reporting is a pervasive shortcoming in program evaluation (Davidson, 2007; Evergreen, 2011), perhaps mainly because reporting is seen to be of less importance than planning, data collection, and analysis. Another reason is that effective communication and reporting constitute demanding tasks that often are not well covered in evaluation training programs.

The communication process is neither simple nor always predictable in practice. Often an evaluator should interact with a client group and provide them with formative feedback throughout an evaluation and beyond to assist with making program improvement decisions. Different segments of an audience may require or be entitled to different amounts and levels of information (Scriven, 2007). In controlling the release of information, an evaluator often has to deal effectively and diplomatically with contractual and legal constraints and sometimes has to cope with pressure from media or political interest groups for premature or inappropriate release of findings. Moreover, it can be difficult and disquieting to present a group with evidence of inefficient or failed program execution, and sometimes an evaluator must reveal findings in the presence of dissension between politically oriented factions. The life of an evaluator is not easy, especially when delivering unwelcome findings to a group of contentious stakeholders!

It behooves every evaluator to engage productively in the development of skills needed to effectively communicate evaluation findings. Fundamentally, evaluators must be masters of written and oral communication. They need to make their presentations factual, interesting, and persuasive. Often they must make complex technical procedures understandable to those with little background in evaluation methodology. The necessary communication process may encompass both informal exchanges and formal reports. Employed media may be oral or printed, textual or graphic, presentational or interactive, published or unpublished, and delivered simply or by complex technology. The process may be open and public or restricted to certain approved intended users. Whatever the nature of the audience and employed media, the reasons to communicate evaluation procedures and findings are to secure understanding and appropriate uses and impacts of the evaluation findings.

Beyond preparing and communicating clear reports and helping intended users apply evaluation findings, evaluators must deal effectively with the psychological and political aspects of evaluation as a change process (Patton, 2008). Clearly, sound evaluation is a change process because evaluators aim to convince their clients and other users of evaluation reports to respect, understand, and use evaluation findings to make and implement appropriate decisions, especially those oriented to improving programs. As discussed in Chapter 16 on utilization-focused evaluation, it is a well-known axiom of any change process that meaningful involvement in crafting and understanding change activities inclines participants to support a program of change. Evaluators need to be skilled in appropriately involving program stakeholders to help plan, conduct, and report on evaluations so that stakeholders will understand, respect, and use findings.

In addition to developing psychological support for evaluations (Donaldson, Gooler, & Scriven, 2002), evaluators need skills in regard to the political aspects of evaluation. In many

evaluations, the intended users of findings include multiple groups with different, often conflicting interests. Not infrequently, one group will seek to exploit the evaluation process and findings to gain or maintain an advantage over one or more other groups. Moreover, it is common for stakeholder groups to engage in heated exchanges about the meaning and practical implications of evaluation reports. Evaluators must therefore be even handed and scrupulously ethical in serving the evaluation needs of all persons with the right to receive the findings. Evaluators should not take sides in disputes between stakeholder groups. Also, they should give voice to all interest groups when it is appropriate to gain stakeholder inputs. Often it is wise to engage representatives of different interest groups to critique evaluation plans and draft instruments and reports, discuss with each other their different judgments of the evaluation work, and convey both their agreements and disagreements to evaluators. Evaluators should respectfully consider inputs from all stakeholder groups. While showing due respect for stakeholders' efforts to offer feedback, however, evaluators often can (and should) disagree with inputs that lack logic and merit, and they must not give away their evaluation responsibilities, authority, and independence. Moreover, evaluators should demonstrate that they value and make appropriate use of critical feedback about evaluation plans, draft reports, and other evaluation materials.

To effectively address political threats to an evaluation's success, evaluators especially need skills to organize and chair discussions among program stakeholders. They should be adept at anticipating; forestalling if possible; and, as necessary, managing conflict. If organizational and personnel conflicts exist, these should be recognized as being very much a part of an evaluation and should be recorded in interim and final reports. Such conflicts will impinge on programs and have been known to prevent the achievement of intended outcomes. Reporting of debilitating organizational conflicts is a professional duty of evaluators and should not be evaded.

Establishing Conditions to Foster Use of Findings

An evaluator should not expect that an evaluation's intended users will automatically make appropriate, informed use of the evaluation findings (Patton, 1997, 2003, 2005b, 2008, 2012). Instead, he or she should work with the client to determine the need for evaluation services following submission of a final report. Such planning has implications for budgeting evaluation work beyond the collection and reporting of information, so that the evaluator then will be available to promote and support use of findings. The evaluator should help the client consider areas of postreporting service (Coryn, 2006; Scriven, 2007), which could include interpreting findings pertaining to new questions from the program's stakeholders, identifying the training needs of program staff, helping the client assess whether a new budget sufficiently addresses issues found in the program, increasing public understanding and acceptance of a successful program, or planning for a follow-up investigation to address identified issues. The evaluator can help members of the client group look into the data produced by the evaluation for relevant, valid information pertaining to such matters and can assist in disseminating findings. However, he or she must not assume the role of the client, including making decisions based on evaluation findings.

In preparing for a possible evaluation study, an evaluator should and can do much to promote use of evaluation findings. An especially valuable step is to arrange for the involvement of stakeholders from the start of deliberations to decide whether to undertake an evaluation. The evaluator should subsequently maintain a close association with stakeholders throughout and even after completion of the study. Once the evaluator and the client have decided to proceed with an evaluation, they should negotiate a contract with strong provisions—budgetary and otherwise—for promoting effective, proper use of evaluation findings (see Chapter 21). We have often found it advantageous to establish and obtain the services of a broadly representative stakeholder evaluation review panel throughout an evaluation. Such a panel can review and give feedback on draft evaluation plans, tools, and reports. We think there is no more powerful means to promote use of evaluation findings than involving users in the process of producing those findings, and having users review evaluation materials is an appropriate way to accomplish this. Following are discussions of four important ways to foster use of evaluation findings that often can and should be employed in the process of launching an evaluation.

Involve Stakeholders in the Evaluation

A key principle to follow is that one's involvement in an evaluation can strongly influence one's understanding of, respect for, and use of the results of the evaluation process. Thus, so far as is feasible, it is wise to involve members of the evaluation's audience in reviewing and reacting to evaluation plans, draft reports, and other evaluation materials. A client most likely is but one element of a right-to-know audience. An evaluator should consider a wide range of potentially interested parties and identify all those who need and have the right to receive evaluation reports. He or she should project how the identified intended users could use the evaluation findings beneficially and should engage as many of them as are available and willing in helping to make such determinations. Helping intended users early in the evaluation process to focus on evaluation questions and to consider how they might respond to answers is a good way to promote their interest in the study and their eventual uses of findings. Patton's situational analysis (1997, 2008, 2012) is particularly useful for this purpose.

Determine Intended Users and Their Potential Uses of Evaluation Findings

Implicit in such evaluation user involvement is the important task of identifying intended users of reports. To be effective, a report must target, as one of the user groups, those involved in a program who at least initially are best placed to propose questions for the study and who logically will be most concerned about evaluation outcomes. This strongly aligns with our often-stated contention that any useful evaluation should have an impact on improving a subject program. An evaluation's intended users should also include stakeholders who will use or be affected by a program's implementation and outcomes. Unfortunately, it is common for evaluators to ignore some stakeholders that not only could give clarity and definition to an evaluation but also could act against the best interests of the study's purposes out of understandable chagrin at being left out. Moreover, such potential users, lacking awareness

of relevant insights arising during the evaluation and becoming confused about aspects of reports, may make decisions of their own accord, to the detriment of program development. In defining the audience for an evaluation report (often working in conjunction with the client to do so), there is the ever-present difficulty of pinpointing the most appropriate people. A sound rule of thumb is to assign a high priority to those persons and groups who will use an evaluation's outcomes to strengthen a program and often an organization itself. In this audience identification process, depending on the circumstances, a stakeholder evaluation review panel may be approached to review the evaluator's definition of intended users of evaluation reports to ensure that all pertinent stakeholders are represented.

Table 24.1 provides a framework for identifying intended users of evaluation findings and their intended uses of the findings. The cells in the first column list a wide range of potential users of evaluation findings, and the headings of the remaining columns identify typical uses of findings. In the first use, program leaders and staff apply findings to focus and develop a program, as occurs in the context and input stages of the context, input, process, and product (CIPP) model (see Chapter 13); also, staff members work toward ongoing improvement of the program, as occurs in Scriven's formative role of evaluation and the CIPP model's process stage. In the second use of findings, those responsible for program operations and results compile evaluative information into accountability reports and present these to the program's financial sponsor and other interested parties, such as beneficiaries. The table's third use of findings focuses on groups that may want to adopt the program and apply it elsewhere. In the table's fourth use of findings, various groups, such as program staff, a professional organization, or a professional journal, may want to publish the findings and distribute them broadly. The fifth noted use involves studying findings to become better informed about the involved phenomena. Among the groups who merely want to study the program's findings may be researchers, college and university instructors, graduate students, and other members of the scientific community. Clearly, many evaluations have a diverse audience with varied interests in the findings. Evaluators are advised to analyze the audience carefully at the outset of and throughout a study to provide a basis for communicating findings effectively and receiving feedback of use in making the evaluation a high-quality, respected enterprise.

We suggest that evaluators use this table to make an initial approximation of the evaluation's intended users and how different parties could be expected to use an evaluation's information. This initial approximation could be represented by checkmarks in the appropriate cells. After evaluators interact with program stakeholders to validate and finalize such approximations, they can apply the analysis, along with other information, to decide what reports would make the most impact, what persons and groups should receive which reports, how reports should be tailored to the different parties' interests, and when and how reports should be delivered.

To make such determinations, evaluators should learn as much as possible about an evaluation assignment. When feasible, they should do so before they agree to do a study, complete a design, or negotiate a contract. They should explore with the client the question of appropriate report recipients and their information needs, and should obtain and study relevant documentation, especially the program proposal. Other relevant background materials could include the needs assessment that led to the development of the program, a request for proposal

Table 24.1 Format for Identifying Potential Users of an Evaluation's Findings and Determining How They Will Use the Findings

	Potential Uses					
Potential Users of Evaluation Findings	**Program Development and Improvement**	**Program Accountability**	**Program Adoption**	**Dissemination**	**Understanding of Involved Phenomena**	**Other**
Client (the person who requested the evaluation, such as the head of the organization in which the program is housed)						
Funder (if different from the client)						
Program director						
Program staff						
Policy board						
Program advisory committee						
Program recipients						
Minority group stakeholders						
Persons who might be harmed by the evaluation						
Potential adopters of the program						
Personnel of competitive programs						
Scientific community						
Instructional and training staffs						
Government programs						
Foundations						
Legislators						
General public						
Media outlets						
Libraries and archives						
Stakeholder evaluation review panel						
Legal community						
Other						

that led to the submission of a proposal for funding the program, pertinent correspondence, minutes of meetings, press clippings, and other significant materials. Evaluators should also contact representatives of different segments of the evaluation's audience, ask them to identify issues that they believe should be studied, invite their identification of sources of relevant information, and ask them to identify other persons who should be interviewed.

An evaluator especially should find and interact with all who might be harmed as a consequence of the evaluation. They should be invited to state any concerns they have about the evaluation's potential fairness and what they think should be done to protect the evaluation's integrity.

We emphasize that, if possible, evaluators should perform a background investigation before agreeing to do a study. Such an investigation is invaluable for deciding whether to conduct an evaluation and, if so, how to design, conduct, and report on the study so that it will be fair to all parties, and so that its findings will be used. A preliminary background investigation has special value for identifying all segments of the audience, engaging representatives to help focus the study, convincing them that the study is worth doing, convincing them that the study will not be a witch hunt or whitewash, and determining with their help how they might use the evaluation findings.

Build Trust and Viability Through Evaluation Contracting

Evaluators are wise to negotiate appropriate contractual agreements that safeguard their ability to interact equitably and appropriately with all stakeholders and ensure a study's integrity. Contracts provide a basis for settling disputes about such matters as which groups should receive which reports; who will help edit reports; who will have final editorial authority; how reports will be evaluated and finalized; and how, when, and to whom they will be disseminated. We believe that the evaluator should insist on final editorial authority and secure advance agreements in regard to the right-to-know audience and release of findings.

Negotiating a sound evaluation contract helps set the conditions for effectively disseminating evaluation findings. Such a contract should clearly define the evaluation's right-to-know audience, the evaluation questions, a schedule of interim and final reports, which reports will be provided to each segment of the audience, opportunities that stakeholders will have to contribute to the evaluation, authority over editing and disseminating reports, any provisions for prerelease review of reports, opportunities for program personnel to rebut reports, and provisions for reviewing and updating contractual agreements as needed. Although evaluators will contract with a client, before signing an agreement they should at least consult with representatives of groups that will be directly affected by the subject evaluation. Clearly, reaching and making public an appropriate set of advance evaluation agreements does much to build trust with program stakeholders and can incline them to respect the evaluation and make use of its findings.

As a practical matter, an evaluation contract should provide for financing evaluation services related to promoting and supporting appropriate use of findings following delivery of the final report (Coryn, 2006). In many evaluations, there is only a vague notion that the evaluator will assist users in interpreting and applying evaluation findings after they have been reported. If the evaluation contract and budget do not provide for funding the evaluator's follow-up involvement, he or she will be unlikely to help users understand, interpret, and apply the findings. The lack of such assistance can render an otherwise sound evaluation relatively cost-ineffective, and failure to budget for such follow-up assistance can be penny-wise and

pound-foolish. It is in an evaluation client's interest to anticipate the need for and fund the follow-up services of the evaluator so that there will be a maximum return of evaluation use for the typically much larger investment in collecting and reporting evaluation findings. (For a broader discussion of evaluation contracting, see Chapter 21.)

Establish a Stakeholder Evaluation Review Panel

In proceeding with an evaluation, we have often found it important to appoint and engage the services of a stakeholder evaluation review panel that includes representatives of the different segments of the evaluation's audience, as illustrated in Table 24.1. The panel's role is to review and provide critical reactions to draft versions of the evaluation design, schedules, instruments, reports, dissemination plans, follow-up plans, and the like. The panel should be charged with assessing the draft materials for accuracy, clarity, feasibility, relevance, importance, and likely impact on the subject program's staff and other stakeholders.

We stress that this panel should be labeled a "review panel," not an "advisory panel" or "steering committee." Providing critical reviews is within the capabilities of a wide range of stakeholders. It is also a vitally important formative metaevaluation task, as discussed in Chapter 25. However, such a group typically lacks, within its membership, sufficient capability to suggest how deficiencies in evaluation materials should best be overcome. Also, steering an evaluation is in the evaluator's sphere of responsibility and authority. Evaluators must not give away this role to a stakeholder review panel, as is wrongly advocated in empowerment evaluation, which we consider to be a pseudoevaluation approach (see Chapter 5). We have seen disastrous consequences when an evaluator has delegated to a steering committee the authority essentially to direct the evaluation work. Results of such a transfer of control can be confusion, conflict, and filibustering among members of a heterogeneous group who believe they can and should decide how an evaluation should or should not proceed. Another possible counterproductive effect is role conflict between the evaluator and the steering committee. Members' judgments on how evaluations should be carried out and on what findings and conclusions should and should not be reported are all too easily influenced by their vested interests related to the program. Although evaluators should be open to suggestions from a review panel, they should emphasize that the panel's main role is to critique, not to engineer evaluation activities. Given this stance, evaluators can receive possibly contradictory critiques from different stakeholders with different points of view and assess and use these inputs according to their merit. Of course, evaluators should inform the panel periodically about how plans and evaluation tools may have been modified in response to the panelists' inputs.

Example of a Stakeholder Evaluation Review Panel

In an evaluation of a state's teacher evaluation system, the evaluation team benefited by appointing and using the inputs of a review panel that was broadly representative of groups with interests in the state's education system. The state superintendent of public instruction chaired this panel. Two members of the evaluation team served as the evaluation's leaders. They provided the panel with advance evaluation materials for review, attended panel meetings,

listened to the panel's critiques of evaluation materials, asked questions as appropriate, responded to panelists' questions, prepared reports of each panel meeting, and used the panel's critiques to strengthen the evaluation and promote use of findings.

The panel members were commanding officers of the state's two large military posts (because many children of servicepeople attended the state's public schools), the president of the state teachers' union, an elementary school teacher, a middle school teacher, a high school teacher, an elementary school principal, a middle school principal, a high school principal, the director of the state's teacher evaluation system, a member of the state board of education, the director of the federally supported educational research and development center located in the state, a dean of an area college of education, an educational measurement specialist, a data processing specialist, a representative of the state chamber of commerce, a representative of the state office of the National Association for the Advancement of Colored People, the majority leaders of the state's senate and house of representatives, a representative of one of the state's largest industries, two high school students, a parent with children in an elementary school and a middle school, and a parent of a high school student. This panel was broadly representative of those who could be expected to be interested in seeing the state's teacher evaluation system strengthened, and it included persons with relevant substantive and technical expertise.

The panel met for about ninety minutes approximately every six weeks throughout the eight-month evaluation. About ten days prior to each meeting, the evaluation's director supplied the panelists with materials to be critiqued plus an agenda (prepared with the panel's chair) and asked each panelist to review the materials prior to the meeting. During the meeting, the state superintendent led the panel through the agenda, which consisted mainly of hearing panelists' critical reactions to the subject evaluation materials and their responses to questions posed by the evaluation team. A special responsibility of the chair was to ensure that all panelists had the opportunity to be heard and that no panelist dominated the discussion. An especially effective technique for managing discussions in a large group is to place a tented name placard in front of each participant and ask anyone who wants to speak to set the placard on its end. The chair can then recognize each person who has something to say in the sequence in which placards were turned up.

One evaluation team member kept notes of the meeting. As needed and requested by the chair, the evaluation's director helped keep the meeting focused on its agenda items. Although there was never an attempt to force a consensus, each meeting was concluded by inviting each panelist in turn to state what he or she considered the most important point for consideration by the evaluation team. Group members could choose to pass on this opportunity to speak. Following each meeting, a member of the evaluation team prepared the minutes of the meeting and sent them to the state superintendent of public instruction for his distribution of copies to the review panel's other members.

This use of the evaluation review panel proved to be highly effective. The panel provided valuable critiques and related information to the evaluation team. The team used these inputs to address issues in the evaluation as appropriate. The team also kept the review panel apprised of how their inputs were being used to strengthen the evaluation. In the process

of reviewing evaluation materials, the panelists became familiar with the evaluation process and eventually its findings. They also came to understand and respect the different points of view represented by the panelists. To be sure, issues were raised and debated. From meeting to meeting, however, the panelists found it increasingly easy to reach and express consensus opinions. Without question, members' participation on the panel inclined them to respect evaluation process and findings and support use of the findings for reforming the state's teacher evaluation system. Moreover, through their critiques, panelists contributed to the evaluation's quality and impact. Following delivery of the final report, panelists were also helpful in disseminating findings to their reference groups.

Checklist for Conducting Review Panel Meetings

We have found the checklist that appears in Exhibit 24.1 to be effective for ensuring efficient, evenhanded, professional exchange among a diverse group of stakeholders at review panel meetings. Although this checklist is keyed to securing panel reviews of draft evaluation reports, with slight modification it is also applicable to obtaining reviews of other evaluation materials, such as standards to guide the evaluation, data collection instruments, and plans for disseminating findings.

Exhibit 24.1 CHECKLIST FOR EFFICIENT CONDUCT OF EVALUATION REVIEW PANEL MEETINGS

_____ Work with the client to identify and recruit members for a representative stakeholder review panel.

_____ Engage the client or her or his representative to chair the panel.

_____ Appoint a recorder for review panel sessions.

_____ Schedule review panel meetings to occur about two weeks after drafts of key reports (or other evaluation materials) will have been developed.

_____ About ten weekdays in advance of a review panel meeting, distribute the draft report to be read in advance of the meeting.

_____ At the meeting, place on the conference table in front of each participant a tent-shaped placard with her or his name printed in large letters on both sides.

_____ Inform the participants that when desiring to speak they should set their placard on end so the chair can recognize them.

_____ Make it clear that for the sake of efficient communication, panel members should speak only when called on by the chair.

_____ Define the group's task as being to review the predistributed draft report.

_____ Make it clear that the panel is not expected to offer advice concerning the evaluation's technical approach.

_____ State the session's main objective—that is, to identify the strengths and weaknesses of the draft report, including inaccuracies and ambiguities.

_____ Stipulate that all inputs related to the session's objective are welcome and that everybody is encouraged to offer his or her input.

_____ Confirm that all inputs from participants will be recorded in the session's minutes, distributed to panel members, and systematically considered in correcting or strengthening the evaluation report.

_____ Note that panelists will see the fruits of their inputs in subsequent, finalized reports that will be sent to them.

_____ Engage a member of the evaluation team to brief the review panel on key aspects of the draft report and to identify key questions requiring responses from panelists.

_____ Provide a brief period for participants to pose questions for response by an evaluation team member.

_____ Devote the major part of the meeting to receiving panelists' comments concerning strengths and weaknesses of the predistributed draft report.

_____ Do not strive to resolve disagreements between inputs from participants during the meeting, but accept and record any clarifications panelists want to offer, including their attempts at resolution.

_____ In concluding the meeting, invite each panel member to present a bottom-line statement denoting his or her view of the meeting's most important message, which could be, for example, a judgment of the evaluation report's value, a recommendation for strengthening future evaluation operations, particular information that should be included in the next report, and so forth.

_____ Conclude the meeting with the chair summarizing what he or she heard, thanking all for their participation, and projecting the panel's future involvement.

Providing Interim Evaluative Feedback

Once an evaluation has been contracted and funded, the evaluator conducts the data collection and analysis activities. Throughout this process, the client group often requires interim reports. This is particularly the case when the evaluation is oriented to supporting program development and improvement, often called the formative role. Not all evaluations serve formative roles; with those that are exclusively summative in their orientation, there may be little or no need for interim reports other than progress reports given to the client to show that the evaluation is on track.

In some cases, an interim report is mainly an early approximation of the eventual final report. This is especially so in applying Scriven's consumer-oriented approach to evaluation.

Under this approach, an interim report addresses those bottom-line questions for which data are currently available. Successive interim reports will have gaps where data are not available, but over time the reports will become increasingly complete in addressing all the items in Scriven's Key Evaluation Checklist (2007; see Chapter 14 and www.wmich.edu/evalctr/checklists).

Evaluators may determine the contents of interim reports by employing any of a number of evaluation approaches. For example, in applying the CIPP model, successive reports may be structured to answer questions about context, inputs, processes, and products, as discussed in Chapter 13 and detailed in the CIPP Evaluation Model Checklist (www.wmich.edu/evalctr/checklists).

Beyond addressing the requirements of an employed evaluation approach, evaluators should also consider and, as feasible, be responsive to stakeholders' questions as they emerge. This requires ongoing interactions between evaluators and stakeholders (as in an evaluation review panel approach). Such ongoing interactions are strongly present in the CIPP model, Stake's responsive evaluation approach (see Chapter 15), and Patton's utilization-focused evaluation (see Chapter 16 and his checklist at www.wmich.edu/evalctr/checklists).

In our CIPP model–oriented evaluations, we have developed and applied an interim reporting approach that we call the "feedback workshop technique." This is a method for systematically conveying draft interim findings to a program's leaders and staff, guiding their discussion of the findings, obtaining their critical reactions to draft reports, supporting their use of findings, and using their feedback to strengthen evaluation plans and materials. In applying this procedure, the evaluator sends draft reports (and possibly drafts of other evaluation materials) to a group as jointly determined by the program director and the evaluator. Typically this group comprises key program staff members and may have a broader composition.

The evaluator sends an interim report to the designated group approximately ten days in advance of a feedback workshop, asking the members to review the findings prior to the workshop. The evaluator asks members of this group to identify any factual errors and ambiguities that they see in the report (and other materials that may have been sent). It is appropriate for the evaluator to solicit and consider critiques concerning clarity and defensibility of conclusions. He or she should not, however, invite program staff members or others in attendance to change evaluative conclusions, because these are the evaluator's responsibility. Although the evaluator should listen to critiques of draft conclusions, he or she must not relinquish his or her authority over determining conclusions. Exercising control over the statement of conclusions is essential if the evaluator is to maintain credibility as an independent investigator.

At the workshop, the program director presides and leads the group through an agenda. Basically, it includes the following items:

1. The program director summarizes the workshop agenda and engages those present to finalize and approve it.
2. The evaluator briefs those at the meeting on the draft evaluation report and any other associated materials (using either overhead projector transparencies or a PowerPoint presentation with associated handouts). He or she then invites attendees to identify factual

errors and ambiguities in the draft. He or she notes that as an independent evaluator, he or she has to maintain authority over the evaluation's findings and conclusions and is not asking those present to amend either of these.

3. The program staff members and others present offer critical reactions to the draft evaluation report and other materials that may have been provided. The reactions may be in the form of oral comments, written notes handed to the evaluator, or marginal notes on copies of the draft report or other materials.
4. The program's representatives subsequently discuss the relevance of findings to possible program improvement initiatives.
5. As appropriate, the program director engages the staff and others in formulating program improvement decisions.
6. The program director invites each attendee to identify the most important point that surfaced in the meeting.
7. The evaluator briefs the attendees on upcoming evaluation activities and may request assistance in carrying them out, such as helping with distribution of questionnaires or arranging access to program files.
8. The evaluator invites attendees to identify any new evaluation questions that should be addressed in a subsequent interim report.
9. The evaluator engages the program's representatives in planning and scheduling the next feedback workshop.
10. The program director summarizes and adjourns the meeting.

Immediately after the feedback workshop, the evaluator may meet informally with the program director and some staff members, perhaps over lunch. Experience indicates that parties find such informal exchanges valuable for applying the interim findings to program improvement, strengthening the evaluation, and strengthening communication between the evaluator and the client group.

Following each feedback workshop, the evaluator prepares the minutes of the meeting and sends this document to the program director for distribution to staff and others as appropriate. The evaluator also uses information from the workshop to correct any factual errors and ambiguities in the draft evaluation report and other evaluation materials that may have been critiqued. He or she finalizes the report and accompanying materials and sends the updated versions to the program director for distribution, as appropriate based on advance agreements.

The feedback workshop technique is keyed to effecting two-way communication about interim findings between evaluator, program staff, and possibly other members of the client group. The technique has proved effective for keeping interim feedback focused on program improvement needs and for helping client groups make relatively immediate use of findings for program improvement. The technique also is invaluable for providing the evaluator with critical feedback of use in strengthening draft reports and other materials. These exchanges have value for keeping interim feedback relevant to emergent program developments, updating evaluation plans as appropriate, and obtaining the assistance of program staff in subsequent data collection

activities. This technique has proved so useful that we advocate its use in any evaluation that involves the provision of interim feedback. (A checklist by Gullickson and Stufflebeam [2001] for applying this technique is available at www.wmich.edu/evalctr/checklists.)

Preparing and Delivering the Final Report

Most evaluations culminate in a final report. Multiyear evaluations typically require annual evaluation reports as well as the ultimate final report. Basically, the annual or final reports should be comprehensive yet reader friendly. This often means that the evaluator should provide different versions of his or her final report—for example, an executive summary; a full report of findings and conclusions; and an extensive technical appendix (or separate technical report) detailing procedures, tools, and data. Final reports may be printed or posted on a Web site. They may be supplemented with audiovisual materials for use in presenting and discussing the findings in group settings.

The example report outlines we provide in this section should be useful for evaluators as they consider how to organize their own culminating reports. However, we emphasize that reporting needs of different evaluations will vary and that no one outline is sufficient for structuring all evaluation reports. Instead, evaluators should carefully study the needs of their audiences and exercise responsiveness and creativity in crafting reports.

Formats for Final Evaluation Reports

Formats for culminating reports vary according to the evaluation approach being followed, the requirements of the evaluation's sponsor, the nature of the evaluand, the homogeneous or heterogeneous nature of the audience, and stakeholders' particular evaluation questions and special information needs. Some reports assess the relative merits of an array of alternative programs, services, or products. Other reports will focus on the background, structure, costs, implementation, and outcomes of a single program or other object. Whatever the report format, it should conform to advance agreements on reporting.

Consumer Reports

Consumer reports that assess alternative objects focus on classes of objects that are available to consumers, alternatives within each class, and the relative costs and merits of objects within each class. *Consumer Reports* is the gold standard for reports that inform consumers about the relative merits of alternative products and services. Any issue of that magazine provides excellent examples of such reports. Within this book, Chapter 10 is an example of a consumer-oriented evaluation report because it critically examines the relative merits of nine alternative evaluation approaches against thirty standards for evaluations.

Another example appears in Exhibit 24.2. It is an outline for a hypothetical evaluation report focused on assessing alternative computers for classroom use. As this outline illustrates, consumer-oriented evaluation reports are keyed directly and concisely to serving consumers' decision-making needs. These reports focus on a particular consumer need, such as to select a type and brand of computer for use in classrooms. In drafting such a report, an evaluator

arrays a reasonable range of decision alternatives—as determined in the evaluation's original design—and classifies them into types (for example, laptops versus desktops, Macintosh versus Windows based).

Exhibit 24.2 ASSESSING COMPUTERS FOR CLASSROOM USE

Contents

- Memory and RAM
- Advantages
- Disadvantages
- Overall merit

Tabular comparisons of five low-budget desktops on . . .

- Cost
- Six technical criteria
- Memory and RAM
- Advantages
- Disadvantages
- Overall merit

Tabular comparisons of five moderate- to high-budget desktops on . . .

- Cost
- Six technical criteria
- Memory and RAM
- Advantages
- Disadvantages
- Overall merit

Best buys

- For classroom use with a low-budget laptop
- For classroom use with a moderate- to high-budget laptop
- For classroom use with a low-budget desktop
- For classroom use with a moderate- to high-budget desktop

The methods behind the ratings

Clearly, a consumer report format is useful when assisting the report's users to critically contrast and compare alternatives to reach an adoption or purchasing decision. Often such alternatives are in the form of products, such as vacuum cleaners, refrigerators, or textbooks. The alternatives can also be softer, as in different approaches to delivering reading instruction or, as seen in Chapter 10, different approaches to evaluating programs. Clearly an evaluator's repertoire for conveying evaluation findings should include the consumer report approach.

Single-Object Reports

In contrast to consumer-oriented evaluations, many evaluations focus on a single program or other object. Rather than helping audience members in making adoption or purchasing decisions, the final reports of these studies typically sum up the nature and accomplishments of the given object. Decisions to be served may include deciding to continue, reduce, expand, improve, or terminate a program. These final reports often are keyed to informing a broad

audience about the program's background, structure, implementation, costs, main effects, and side effects.

Exhibit 24.3 contains the contents page for a summative evaluation that focused on a single project—the self-help housing project referenced in previous chapters. The final report of this evaluation is noteworthy for its unique organization. Like most other final evaluation reports, all of the contents of this report (except for a detailed technical report) were inside one pair of covers. However, this report essentially was three reports in one. The intention behind using this format was to provide different reports for study by different segments of the audience and to make it easy for members of each segment to locate quickly and read the part of the presentation that most interested them.

Exhibit 24.3 CONTENTS PAGE FOR THE SELF-HELP HOUSING EVALUATION

Contents

10. Evaluation Findings (context, inputs, process, impact, effectiveness, sustainability, transportability)
11. Conclusions (project strengths, project weaknesses, key lessons learned, bottom-line assessment)

(Photographic Reprise)

Epilogue

Acknowledgments

References

About the Evaluators

Appendices

A. Evaluation Reports
B. Traveling Observer's Handbook
C. Case Study Participants
D. Case Study Interview Protocol
E. Builder Interview Protocol
F. Evaluation Personnel
G. Metaevaluation (attestation of the evaluation's adherence to professional evaluation standards)
H. CIPP Evaluation Model Checklist
I. Evaluation Contract

Another unique feature of this composite report was that each individual report concluded with a photographic retelling of the report's story. This was feasible in this evaluation because of the visible nature of a host of vital project elements. Photographs pertaining to Report One on project antecedents showed the beauty of nearby mountains and ocean beaches; the run-down nature of area neighborhoods; the presence of squatters on area beaches; a village of small, two-room structures for homeless people; area schools and recreation facilities; area stores and restaurants, including boarded-up buildings; leaders of the project's sponsoring foundation; and project staff in their planning sessions at foundation headquarters. Photographs recapping the Report Two story of project implementation portrayed project staff; vacant building plots prior to construction; contractors providing project beneficiaries with on-the-job instruction in construction procedures; beneficiaries at work building their houses, including women and men operating power equipment with which they were previously unfamiliar; the project's

community center and playground; the variety of types of houses built; houses in various stages of construction; cul-de-sacs designed to foster community cohesion; landscaping around houses and the community at large; expensive stone walls and heavy iron gates that home owners had erected to mark off their property and keep it secure; the celebration and blessing that followed each construction cycle; happy families each inside or outside their just-completed house; and participation of children and their parents in educational and social support activities. Photographs reprising the Report Three story of the project's results confirmed that after seven years, the foundation and beneficiaries had produced a wonderful community made up of attractive houses and well-kept yards plus a diverse population of grateful, low-income home owners and their families.

The photos showed happy children—in front of their well-constructed houses, at play, and being mentored by a local teacher in the community's study center; nicely landscaped properties; happy couples proudly standing in front of the house each had built; community picnics and parties; an impressive center designed to house community activities; and a large number of home owners in a community meeting. This set of photos was concluded with two adjacent photos intended to sum up what had been accomplished. One showed the project's benefactor, Consuelo Zobel Alger, who prior to her death had declared, "I want to spend my heaven doing good on earth." The other concluding picture was a panoramic view of the beautiful, child-friendly new community, named the Spirit of Consuelo.

In this evaluation, the sequence of photographs at the end of each report had a number of advantages for readers. It helped them better assess and appreciate why and how the project was started, where it was conducted, whom it served, who led the project, who performed the various tasks, how the project was designed and carried out, how it progressed over time, and what it accomplished. As the evaluation's director, this book's first-named author observed that reviewing photographs stimulated some readers to return to certain sections of the report to deepen their understanding of project implementation and accomplishments. The pictures also stimulated and aided discussion about certain issues in the project, such as the values underlying the project and home owners' propensity to mark off their respective properties with impressive stone walls and gates. Some readers of the report told us they would have been skeptical about its printed claims of high project success had they not seen the successes vividly displayed in photographs. Based on our experience in this evaluation, we strongly recommend the use of photographic reprises when a project's characteristics are highly visible, and especially when a project's progression and success (or failure) can be demonstrated in an appropriate sequence of pictures. Now we look more closely at the composition of the Spirit of Consuelo evaluation report (available from the book's Web site at www.josseybass.com/go/evalmodels).

Report One was aimed at potentially interested persons who had no prior knowledge of the project, its sponsor, or its environment. This report presented information on the Consuelo Foundation, the values it imparts to its projects, why and how this organization started the self-help housing project; the assessed needs of working poor and homeless people in the area; the project's environment of mountains, seacoast, and poverty-stricken neighborhoods; and its economic, demographic, and social characteristics. This report addressed each of these matters in some detail. Persons from around the world who had expressed interest in this

project wanted such background information. The evaluation's client group was already well informed on these matters, however, and needed only to skim the contents of Report One.

Report Two presented the details of project implementation. It was especially addressed to persons and groups that might be interested in replicating the project's unique approach to values-oriented, self-help housing and community development. This report presented an overview of the project's rationale, values, targeted beneficiaries, goals, structure, staff, operations, and financing. Especially, this overview explained how, during each of seven annual increments, six to seventeen pairs of cobuilders had worked together on weekends over a ten-month period to build the houses that on completion would be assigned by lottery to the builder pairs. The report presented the criteria for choosing project participants and explained how they were recruited and chosen. It described the selected cobuilders, including their backgrounds, demographic characteristics, ethnic composition, education levels, areas of employment, and especially their children. It explained the approach to helping project participants obtain mortgages and related loans. It described the foundation's investment in the project and how it would recoup some of the investment. It explained how project participants built their own houses, how they were trained and assisted to do so, and how they were supervised. It described the roles of the contractors who trained the project participants and of those who did the specialized electrical, plumbing, and concrete work. It described the state's role in constructing and maintaining roads and infrastructure and inspecting the various stages of construction. It reported on why and how the foundation erected a community center and explained the covenants and rules for living in this community. Report Two also discussed how the Consuelo Foundation's staff delivered social and community development services throughout the life of the project. Clearly, this second report was intended to inform interested parties about the project's nuts and bolts and what would be required to mount and conduct similar projects.

Report Three presented the project's results. These were assumed to be of interest to all segments of the evaluation's audience. This report summarized the details of the employed evaluation approach, including the initial grant to plan the evaluation; the evaluation's intended users, their information needs, and the reports that were tailored for use by different groups; employment of the feedback workshop technique; intended uses of evaluation findings; value bases for judging the project; data collection tools and activities; plans for synthesizing and reporting evaluative information; evaluation personnel and costs; constraints on the evaluation, such as prohibitions against recording interviews with beneficiaries; and provisions for evaluating the evaluation. The core of Report Three included findings divided according to the different components of the CIPP model and focused especially on the assessed needs of targeted beneficiaries and the local area. The findings were presented in terms of the project's context, structure, process, reach to the targeted beneficiaries, effectiveness in addressing beneficiaries' needs, sustainability, and transportability. Report Three's ensuing conclusions were divided into the project's strengths and weaknesses, key lessons learned, and a bottom-line assessment. The last focused on the project's quality; worth to beneficiaries, the local area, and community developers around the world; and unfulfilled needs and objectives.

Following presentation of the three core reports, the overall document concluded with an epilogue, recognition of those who contributed to the evaluation, references to the interim

reports that led to the final document and relevant publications, information about the evaluators and their home organization, a pull-out copy of the executive summary, and appendices. The appendices contained data collection instruments, interim evaluation reports, information about project beneficiaries and evaluation staff, a self-assessment metaevaluation by the staff, and a copy of the CIPP Evaluation Model Checklist. The metaevaluation was keyed to the thirty 1994 Joint Committee program evaluation standards. The appendices provided a modicum of information about the design, tools, and implementation of the evaluation that probably satisfied the interests of most members of the audience. Technical specialists in the evaluation community, however, probably would require much more detail about the evaluation's methodology than could reasonably be contained in one document designed to serve all segments of the evaluation's diverse audience. Therefore, in evaluations such as the one described here, we advise evaluators to prepare a separate, detailed technical report on how the evaluation was designed, instrumented, carried out, and assessed.

The Role of Visual Processing Theory in Reporting Evaluations

Only recently has visual processing theory, coupled with significant advances in data visualization, been recognized as a means for more effectively communicating evaluation results. As Evergreen (2010, 2011) noted:

> There is now a somewhat clearer idea of how much visual science and graphic design have been incorporated into evaluation communication and reporting. According to visual processing theory, evaluation report authors are missing opportunities to more fully engage their readers. The use of color, placement, and size to emphasize critical information could help evaluators more efficiently communicate. Some factors, like choice in typeface and color of type, that have the ability to impair legibility appear to be well-managed. Yet, in some areas, authors are designing reports that actively work against reader comprehension. (2011, p. 70)

The statistician John Tukey was one of the first to promote graphic methods for displaying statistical data (for example, he introduced the boxplot in his 1977 book *Exploratory Data Analysis*). Tukey's important work in pioneering data visualization set the stage for the many advances in data display and visualization that have occurred in the last few decades. One of the most influential and far-reaching scholars in data display and visualization in the last quarter century is Edward Tufte. In his book *Envisioning Information,* Tufte (1990) emphasized creating data displays that maintain integrity and even beauty, acknowledging the utility of typography, grid systems, and asymmetrical page layout for making findings vivid and understandable. Referring to an overall report, not simply the visual display of data, he readily acknowledged that the credibility of data can be lost with poor design (Tufte, 1990, 1997). Generally, Tufte (2001) has advocated the following principles for visual representations of data:

- Use color selectively to highlight.
- Use light-grey gridlines (when gridlines are used at all) so as not to distract from the data.
- Avoid line patterns or textures that cause visual activation or optical illusions.
- Ensure that comparisons between text and a graphic or between two graphics occur within the same eye span on a page.
- Make all visual distinctions as subtle as possible, but still clear and effective.
- Make sure that the representation of numbers, as physically measured on the surface of a graphic itself, is directly proportional to the numerical quantities represented.
- Eliminate ink that does not express information.

More recently, Few (2009) focused on the role of color in design, particularly in regard to charts and graphs. As with earlier developments in graphic displays of data, improvements in technology propelled Few's work. He discussed the need to manipulate background colors to contrast appropriately with areas of emphasis in the foreground, even introducing notions of hue and saturation into the working knowledge of applied statisticians. Like others before him, Few insisted that graphic displays of information support the proper interpretation of that information. In his 2011 book, *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*, Yau provided detailed instruction on how to create high-quality statistical graphics in R; how to design in Illustrator; and how to create interactive graphics in JavaScript, Flash, and Actionscript.

Torres, Preskill, and Piontek (2005) authored one of the few books on evaluation reporting and communication, underscoring their advice with a discussion of individual adult learning best practices and of the relationship between report layout and comprehension. But Evergreen (2011), referring to this book, wrote:

> While many graphic design best practices are included, the authors include practices such as pie charts (the first error) with three-dimensional perspective (the second error) and suggestions and examples that illustrate improper design. These include the use of gradated color, thick grid lines, textured charts, and placing graphs in the report appendix . . . Still, to their credit, the authors list graphic design among other skill sets evaluators need for good communication and reporting. (p. 27)

In 2011 Evergreen, as part of her doctoral work at Western Michigan University, developed the Evaluation Report Layout Checklist (see Exhibit 24.4), which is intended to be used as a diagnostic guide to identify elements of evaluation reports that could be enhanced using graphic design best practices and/or the assistance of a graphic design expert. In applying the checklist, one should rate each aspect of an evaluation report (as shown in the "Indicator" column in the checklist) according to the following scale: F = fully met, P = partially met, or N = not met. (The "Best Practice" column can be used as a guide for improvement.)

Exhibit 24.4 EVALUATION REPORT LAYOUT CHECKLIST

Category	Indicator	Best Practice	Notes
Type	Text fonts are used for narrative text	Use serif fonts. Nothing with lots of graphic detail.	Nice *serif choices* include Garamond, Palatino, Cambria Nice *sans serif choices* are Trebuchet, Verdana, Calibri *Sentence case* is when the first letter of the line is capitalized and all others are lowercase, excepting proper nouns. *Body text* is that which comprises the narrative of the report. By contrast, *header text* is that which comprises your headlines and titles. Also known as display text. • Default bullet size (too big) • Appropriate bullet size
	Long reading is in 9–11 point size	Studies have shown that 11 point text is easiest to read at length, but it can depend on the typeface (font).	
	Body text has stylistic uniformity	Each text section has unbolded, normal text in sentence case (no all caps), except in short areas of intentional emphasis. This supports undistracted reading.	
	Line spacing is 11–13 points	For lines within paragraph, generally choose 1–2 points larger than the size of the body text.	

Category	Indicator	Best Practice	Notes
	Headers & callouts are emphasized	Header should be 150%–200% of body text size. Sans serif or decorative is okay. Use sentence case. Contrast with body text by using different size, style, and/or color. Too similar looks unintentional.	
	No more than 3 fonts are used	A change in font will indicate a change in meaning. Use font changes to guide reader through information according to importance.	
	Bullets are slightly less thick than text	If bullets must be used, decrease their size to slightly less (70–80%) than the point size of the font. Otherwise, they are too strong and distracting. If good spacing is used in lieu of bullets, this best practice is Fully Met.	
Alignment	Alignment is consistent	Alignment is a preattentive feature easily picked up by a reader, so be sure elements start in the same place on each page unless misaligned on purpose. Avoid centered elements.	Imagine each page divided into rows and columns. Draw imaginary lines to check that elements are aligned at the start of each row and top of each column. *Asymmetry* is an easy way to create interest. Try placing a cool picture off to one side of the page. *Wide margins* are a quick way to create empty area and manage line length.

(*continued*)

Category	Indicator	Best Practice	Notes
	Columns are 8–12 words in length	This is 50–80 characters, depending on font. Longer is difficult to track from line to line, shorter creates too many hyphenated words, distracting the reader.	
	Important elements are prominent	Most prominent position is top half of page and/or emphasized by size, color, orientation, etc. Supportive information is toned down.	
	Body text is left or full justified	Ragged right edge is more informal, but easier to read for average readers. Full justification is formal, easier for fluent readers, but creates design issues with "white rivers" or large gaps of white space between words.	
	Grouped items logically belong together	Grouped items are interpreted as one chunk. Place logical items together. Add space between groups. Minimize space between header and body text.	
	Empty area is allocated on each page	Leave plenty of space between paragraphs, around page margins, and between text and graphics. It gives eyes a rest.	

Category	Indicator	Best Practice	Notes
Graphics	Pictures/graphic elements are present	Multimode learning increases chance at storage of info in long-term memory because it eases cognitive load of body text. Choose pictures or graphics related to your topic. Graphics include, but shouldn't be limited to, tables and charts. If there are no graphics, this section is all Not Met.	*Pictures and graphics* related to your content will make your content more memorable. *Choose pictures* from quality sources, like paid websites. Watermarks or fuzzy images are signs of an amateur. *Use a cover page* at the beginning of a report. This is a good place for a very large graphic.
	Graphics are near associated text	If readers must flip around to interpret between text and graphic, comprehension will be impaired.	
	Graphics are simple	Less visual noise leads to better assimilation. Eliminate gradation, textures, or graphics as backgrounds. Segment complex graphics into smaller chunks.	
	Size corresponds to changes in meaning	Use, for example, larger pictures on chapter start pages. In graphing, for example, be sure height of columns proportionately represents data.	
	Graphics direct toward text	Use the power of an image to direct the reader's gaze from the image to the associated text. Eyes in a photo, for example, should look inward at text.	

(*continued*)

Category	Indicator	Best Practice	Notes
	Visual theme is evident	Pick a visual theme that can be used in different forms throughout report to give strong emotional connection.	
	Some elements are repeated	Repetition of some graphic elements adds unity to the piece, makes work more memorable. Careful not to overdo it—too many elements can add clutter or complication.	
Color	Narrative text is dark grey or black	Black has highest comprehension levels, with low intensity colors taking a distant second place.	Keep in mind various culture-laden *color connotations*. For example, pink is highly associated with feminine qualities in the United States. Make sure your color choices are appropriate for your audience. Note that *people with colorblindness* have difficulty with red-green and yellow-blue combinations. A safe bet is to use your *client's colors*.
	Background has white/subdued color	Reversed-out text (e.g., white text on black background) impairs information retention.	
	One or two emphasis colors are used	Subdued colors that still contrast with background should be used. When used, it should be to actually emphasize important information, like data in a graph. If more than one is selected, consider choosing along a color gradation so that order of importance is implicit.	

Category	Indicator	Best Practice	Notes
	Color changes mark meaning changes	Color changes signal a change in hierarchy of information. Be intentional with color changes so that a viewer doesn't get confused.	
	Color reprints legibly in black and white	Color looks different on a computer screen than on paper. Print on a black-and-white printer and then make a copy of that printout to check legibility.	

Source: Evergreen, S.D.H. (2011). *Death by boredom: The role of visual processing theory in written evaluation communication*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Two caveats are in order concerning the use of advances in formatting and the use of color in preparing evaluation reports. The first concerns the cultural context in which the reporting is to occur. Organizations in some contexts in which we have worked would probably find highly elegant, colorized, extravagant, technology-dominated reporting to be off-putting. The U.S. Marine Corps, for example, would find such an approach highly alien compared to their customary use of plain briefing sheets and overhead transparency presentations. In working to stimulate interest and secure impacts it can be highly important to report in ways that are familiar to the target audience. The second caveat relates to colorizing reports. Recipients often want to duplicate and distribute such reports, and to do so in black and white. If the original report is in color, black and white reproductions may turn out to be almost illegible, and many client groups would prefer to avoid going to the trouble and expense of reproducing the report in color. Overall, in deciding how best to convey findings, evaluators must consider the usual communication preferences and practices of their client groups as well as technological advances in reporting.

Presentation of Evaluation Reports: Decisions About Conclusions and Recommendations

Final reports may be presented in a number of ways and forms. Above all, what is presented must be clear, concise, and of interest to relevant program stakeholders; otherwise, much of the evaluation's purpose will be lost. The stakeholders' interpretation of outcomes is vital to

the presentation. To achieve sound interpretation and strong desire on the part of stakeholders to use the evaluation's findings, careful thought must be given to planning the form or forms of presentation. Whatever these are, they should attract the reader's interest. Evaluation judgments, the formulation of which constitutes an essential part of the evaluator's task, should directly address the evaluation's questions and be grounded in its data. If the advice given in this chapter is followed, there should be no unjustified surprises arising from conclusions and their presentation. In summary, presentation of findings and evaluator judgments and conclusions must be well organized, logically developed on the basis of the evaluation's procedures and data, and so convincing that the users will readily embrace and use them for program improvement and accountability.

Readers may have noted that we made no mention of recommendations in our depiction of the Spirit of Consuelo evaluation. Evaluation practitioners and theorists differ markedly in their opinions about the place and worth of recommendations. In the past, almost all evaluations concluded with a series of these. Indeed, clients expected to receive recommendations offering guidelines for future decision making. Provided that recommendations arise from logically developed evaluator judgments based on sound information, warranted assumptions, and grounding in widely accepted standards, it would be acceptable for evaluation reports to include advisory recommendations. But caution is needed.

Let us consider the wider implications of evaluation recommendations. Will these typically include adequate safeguards against acting on poor advice? The answer to this question is no. As Patton (1997) stated, "Recommendations have long struck me as the weakest part of the evaluation" (p. 328). He expressed the opinion that despite huge strides in studying programs and in methodological development, evaluation has lagged behind in the formulation of recommendations. Moving from data gathering and analysis to making recommendations is not, to Patton's way of thinking, "a simple, linear process" (p. 328). Readers and users of evaluation reports have every right to know and understand on what bases recommendations were compiled. Scriven (1993) clearly warned that it is potentially fallacious to draw a nexus between determining the merit, worth, and value of an object and making recommendations. He contended that although good evaluators may develop sound judgments, too often they are ill equipped and lack the necessary experience and data to formulate defensible recommendations.

There are, however, other views about the presentation of recommendations. For instance, Hendricks and Handley (1990) have stated their opinion that "evaluators should almost always offer recommendations" (p. 110). Our view is that a responsible evaluator must consider the possibility of including recommendations. The decision to provide recommendations or not should be discussed with the client and other stakeholders at the early planning stage. A decision to consider recommendations must always be qualified by caveats. For example, the study should proceed only to the point of evaluator judgments, not to recommendations, if the obtained data pertain exclusively to the program's merit and worth. In such a case the evaluator appropriately would leave it to the client and other stakeholders to decide how best to respond to the evaluation's findings. If, however, the evaluation includes a systematic follow-up study to identify and assess alternative responses to the evaluation's conclusions, then the evaluator appropriately can identify and recommend a sound course of action. This

is the approach offered in Stufflebeam's CIPP model (see Chapter 13), wherein the need for recommendations in response to an evaluation's assessment of a program's merit and worth is addressed appropriately, substantively, and systematically through a follow-up input evaluation, in which the evaluator identifies and carefully assesses alternative possible follow-up actions.

Providing Follow-Up Support to Enhance an Evaluation's Impact

The completion and submission of evaluation reports lay the foundation for subsequent efforts to apply the findings. Recipients of evaluation reports need to study, assess, and soundly interpret and apply the findings, especially those in annual and final reports. Provided that the advance evaluation agreement includes follow-up functions for the evaluator and resources to support the work, an evaluator can help stakeholders understand and consider how they might appropriately use reports. To the extent that the evaluator is able to do so, his or her follow-up efforts can increase the evaluation's impact.

We reiterate that evaluators should not in any way take over the decision-making role of their clients. It is proper for evaluators to promote and facilitate appropriate use of evaluation findings, but it is inappropriate for them to make program decisions based on the findings or to exert undue pressure for certain choices. Appropriate ways evaluators can foster proper use of findings are to (1) help users interpret findings and see their implications for decision making and action; (2) help them avoid misinterpreting findings; (3) caution against making inappropriate inferences and pursuing inappropriate applications; (4) help potentially contentious groups deliberate civilly and constructively about the findings; and, as appropriate, (5) help stakeholders plan for needed follow-up investigations.

Leading Discussions and Managing Conflict

Dissension about how findings can and should be applied often occurs when an evaluator meets with stakeholders to encourage them to make use of the final report and to assist them in doing so. Consequently, evaluators need to be adept at coordinating group discussions and also at managing conflict when it arises in feedback sessions. This is especially important when, in the course of presenting and leading discussions of evaluation findings, there are heated exchanges over the meaning and importance of those findings. Among the critical skills of conflict management are those involved in engaging "combatant" groups to discuss and resolve their disagreements about what evaluation reports mean and how to use them appropriately. As a discussion leader, the evaluator needs to be able to

- State the predefined goals for group discussion, engage the group in any needed clarification and modification of goals, and solidify agreement on goals for group discussion
- Present the predefined agenda and rules for group interactions, and engage the group to clarify and improve the agenda as appropriate
- Coordinate or assist with coordination of exchanges among group members
- Initially foster divergent thinking and deliberation

- Ensure that all participants have opportunities for input
- Prevent overzealous participants from dominating the exchange
- Listen and take notes
- Ask clarifying questions
- Summarize areas of agreement and disagreement as they emerge
- Help the group try for consensus
- Engage all participants in summarizing what they see as the outcomes of the discussion
- Engage the group to decide on next steps
- Prepare and distribute a summary of the meeting

As is evident in this list of discussion leadership tasks, an evaluator can play a valuable mediation role in helping a diverse stakeholder group understand and appropriately use evaluation findings. Every evaluator should thus develop his or her ability to be an effective discussion leader. The checklist in Exhibit 24.1 provides a convenient list of key checkpoints to consider when planning to lead a stakeholder group's deliberations about evaluation standards, plans, procedures, tools, reports, and so on.

Helping a Client Group Interpret and Apply Evaluation Findings

In our experience, an evaluator's follow-up activities typically are conducted according to agreements reached with the subject program's director. Together, the evaluator and program director define those persons who probably will need follow-up support, such as program staff, a program advisory panel, beneficiaries, interested community members, the program's policy board, and others. Separate meetings may be slated for each of several key groups. It can be important to meet with such groups separately, because program-related responsibilities and interests may vary considerably from group to group. To foster evaluation use, evaluators should tailor follow-up exchanges concerning evaluation reports to the program-related responsibilities of each group and the group's particular interests and questions.

In preparing for and conducting follow-up meetings, evaluators can meet with evaluation review panels and conduct feedback workshops with program staff and others. We suggest that evaluators and their clients will find Gullickson and Stufflebeam's Feedback Workshop Checklist (2001) to be highly useful (available at www.wmich.edu/evalctr/checklists), along with the checklists provided earlier in this chapter. Other approaches to providing follow-up assistance for interpreting and using evaluation reports may include focus groups, public forums, sociodramas, journal articles, Web sites, and simulation exercises based on evaluation findings.

Sociodrama Example of Evaluation Follow-Up Assistance

A very interesting follow-up approach observed by this book's first-named author was a sociodrama designed and conducted by Robert Stake. He had directed an evaluation of a state's educational testing and accountability system and issued a comprehensive, largely

critical report of the system's procedures and findings. The lengthy technical report was of interest and use to a narrow group, including in particular a state education department's officials and accountability specialists. However, a much broader audience of educators, policymakers, and citizens was interested in learning and deliberating about the involved policy issues without having to study the detailed technical report.

To respond to the needs of this diverse group, Stake conducted a series of public meetings in large auditoriums. Each person in attendance was provided with a single sheet that, in matrix form, identified the study's conclusions about the state's new accountability system. Conclusions presented on the sheet included strengths and weaknesses of the accountability system and issues of concern at local as well as state levels.

The riveting factor at each meeting was a sociodrama enacted on the auditorium's stage. Five chairs were arranged as if they were the seats in an automobile. Persons sitting in the "automobile" seats represented a group that had just left a meeting at which the evaluation of the educational accountability system had been presented. The five stakeholders—a teacher, a school principal, a school board member, a member of a local parent-teacher association, and the school district's director of testing—were returning to their home city on the other side of the state.

During their journey, these persons reflected on and discussed what they had heard at the meeting. They mainly talked about issues evident in the accountability system that they judged especially relevant to their school system and community. One topic was the fairness and quality of the state's tests. Although they acknowledged that a reputable testing company had produced the tests, they also questioned whether the company had been given sufficient time and resources to validate the tests. In fact, the testing company had issued strong caveats in this regard. The discussants thought the pass-fail standards attached to the tests probably were arbitrary and unrealistic. They were highly skeptical about the use of scores from a single test to promote and graduate students and to reward or sanction schools. They projected that this high-stakes testing would generate unjust decisions about students, teachers, and administrators and would be likely to create much conflict. They also acknowledged that there were some advantages of employing the same test in every district in the state, especially that a common state testing program would be welcomed by the state's lay public. But they worried that a single set of state tests might have an adverse, narrowing effect on their local curriculum, especially because the tests did not cover all curricular areas. Another worry was that teachers and students would spend inordinate amounts of time practicing and preparing for the state test. Moreover, they worried that this testing program could spawn possible cheating by teachers, who might feel pressure to make sure that all students' test scores were in the acceptable range. Still another concern was that the state would take so long in scoring test answer sheets that results would be returned too late to be of much use to teachers and students. Thinking about the state level, the discussants wondered whether this new testing program would sap the state's scarce resources for education, or, if not, whether the state would be able to revise the tests regularly and keep them up to date.

The actors raised a number of provocative issues. Following the dramatization, Stake added some comments and briefed the audience on the summary of program strengths, weaknesses,

and issues on the single sheet that had been distributed. He then opened the public meeting to comments, questions, and discussion. As it invariably does, this sociodrama format spawned a substantive, provocative exchange among those in attendance.

The preceding reporting approach employed by Stake illustrates how evaluators can be creative and provocative in engaging a diverse audience to understand and consider the implications of an evaluation's findings. We believe evaluators can and should increase impacts of their evaluations by exercising creativity in reporting evaluation findings and engaging the report's intended users to learn about, understand, assess, and act on those findings.

Helping Policy Groups and Program Administrators Understand and Apply Findings

Evaluators often have to serve the information needs of policy bodies as well as program administrators. In serving a range of different intended users, evaluators need to be sensitive and appropriately responsive to a number of issues. Especially, they should be mindful of and attentive to the needs of the client with whom they have contracted. Typically this client is the program's director, but the client might be the program's policy board. In either case, the evaluator and client need to plan together how to contact and serve the evaluation's different users.

In such planning, evaluators should be mindful and respectful of differences in the evaluation needs of policymakers and administrators. Evaluators should not fall into the situation of aiding and abetting policymakers in usurping program administrators' authority. The potential for misuse of evaluation findings emerges when policy boards inappropriately delve into the area of administrative decision making and want evaluators to give them information of use in directing the day-to-day program operations and, possibly, making personnel decisions. Evaluators should not steer policymaking groups in the direction of taking over administrative decisions. Also, an evaluator should not become party to a policy board's decision to hire or fire a program's director or other staff members. The guiding principles here are that program staffs should have authority that is commensurate with their areas of responsibility and that evaluators should provide evaluative information and follow-up services to policy boards and program staffs that are commensurate with their areas of policy responsibility and authority.

Clearly, evaluators have to walk a fine line and exercise utmost professional integrity when serving both policy groups and program administrators. Ideally, policymakers set program goals and priorities, assign responsibility to administrators for carrying out a program, and give them commensurate authority to make and apply the day-to-day program implementation decisions. Although evaluators have no authority to differentiate between policy and administrative roles pertaining to a given program, they can and should be sensitive to potential problems in this area and strive not to exacerbate them. They should avoid focusing a policy board's attention on a program's day-to-day administrative matters. In general, it is appropriate to focus the attention of such a policy group on needs assessment data of use in affirming or revising program goals and priorities, cost-effectiveness information of use in assessing the adequacy

of program resources and the return on investment, process information of use in determining whether a program is a high-quality effort that is focused on intended outcomes, and outcome information of use in deciding whether a program is achieving its goals and is worthy of continuation.

Planning and Conducting Follow-Up Input Evaluations

When promoting the use of an evaluation report, it is wise to consider the particular study that generated the report in a broad, improvement-oriented context. Although evaluations often raise more questions than they answer, the unanswered questions can valuably lead to needed follow-up studies. This is especially the case in using the CIPP model (see Chapter 13), whereby an input evaluation might be launched to help the client group solve problems identified in the original evaluation. When a product evaluation identifies issues and problems in a program, the evaluator and client appropriately should consider whether a problem-solving study should be conducted. If so, they are wise to plan and launch an input evaluation aimed at identifying or developing, and then assessing, alternative solution strategies.

Example of Using the CIPP Model to Reform a System

As discussed in previous chapters, the evaluation conducted for the U.S. Marine Corps (USMC) employed an input evaluation to help USMC leaders reform the corps's personnel evaluation system. The initial evaluation found that the existing system for evaluating officers and enlisted personnel was in need of reform. The leaders of the USMC agreed with the evaluator and his team's conclusion that the personnel evaluation system was in need of reform or replacement. The evaluation had not culminated in recommendations for solving the identified problems, however. Clearly, the study had produced no appropriate information base for solving those problems. Any recommendations the evaluators would have made would have been strictly "armchair" and not defensible. An acceptable solution to the problems the evaluators found had to be located outside the bounds of the original study. Accordingly, with the USMC's buy-in, the evaluators conducted a follow-up input evaluation.

In that study the evaluators identified about ten alternative personnel evaluation systems being used in other military services and in business and industry. The evaluators critically examined these against the USMC's twenty-one standards for sound personnel evaluation systems and found that none was acceptable for solving the problems in the current personnel evaluation system. Collectively, however, these different systems had some promising features.

The evaluators subsequently used the results of this initial input evaluation as a basis for creating plans for three new personnel evaluation systems. This was accomplished by engaging three independent teams, each to generate the best possible personnel evaluation system for the USMC. Each team reviewed the results of the previous evaluation of the existing personnel evaluation system plus the information they had gathered about other personnel evaluation systems, including ratings of those systems. One team was tasked with using this information to generate a system that built on and reformed the existing personnel evaluation system.

The second team was tasked with designing a system based on the outside system the evaluators had rated highest. The third team was tasked with exercising as much creativity as possible to design an "out-of-the-box" new system.

The three teams designed these systems as competing approaches to reforming the existing personnel evaluation system, and they took advantage of what the evaluators had learned through assessing the ten alternative systems. The evaluators evaluated the three invented systems against the twenty-one standards and reported the results to the USMC hierarchy. The top generals chose one of the systems for development and implementation. It turned out to be the system that was a reform of the current system and not the one the evaluators rated best overall. The evaluators had, however, rated the plan for reforming the current system highest on feasibility.

The evaluators then provided a plan for constructing and pilot-testing the new system. The plan was patterned after the process and product components of the CIPP model. The USMC proceeded to develop, pilot-test, and install the new personnel evaluation system. We see this experience as a cogent example of how to follow up an initial conclusion-oriented evaluation by designing and conducting a subsequent solution-oriented evaluation.

Summary

This chapter is titled "Communicating Evaluation Findings," but more than that it is about securing impacts of evaluation studies. Without the proper use of its findings, an evaluation can be no more than an academic exercise. The concern for impact should be pervasive throughout the process of planning, budgeting, contracting, conducting, and reporting on an evaluation. Evaluations typically should yield interim feedback as well as annual and final reports. As feasible, evaluators should support client groups' various uses of reports following their delivery. Clearly, the needs and interests of the full range of stakeholders should be considered when promoting use of findings. Involving stakeholders in the evaluation process is a vital way to secure their interest and stimulate them to value and use evaluation reports. In so doing, evaluators need to be sensitive to the psychological, political, policy, and administrative aspects of programs and communication processes, and they must possess the skills needed to attend to these components.

We have presented sample formats for consumer reports and single-object reports. We have warned that evaluators must not take over clients' decision-making role, and we have discussed some of the difficulties in serving policy bodies as well as administrators. We have also issued cautions about the complexities and difficulties inherent in making recommendations. Often follow-up studies, such as input evaluations, are needed to generate defensible, actionable recommendations. We have suggested a range of techniques for promoting evaluation impacts: evaluation review panels, feedback workshops, public forums, Web sites, focus groups, journal articles, and sociodramas. We have provided or referenced several checklists for guiding efforts to communicate evaluation findings, and we have summarized theory on formatting evaluation communication. We have also shown that the CIPP model is focused directly and practically on promoting evaluation impacts.

REVIEW QUESTIONS

1. List at least four challenges to effective communication of evaluation findings, as discussed in this chapter.
2. In general, who should be included in an evaluation's audience, and specifically, what are at least four likely groups to be included?
3. List at least four provisions related to reporting evaluation findings that should be addressed in a contract for a program evaluation.
4. Define the role and composition of a stakeholder evaluation review panel and the process for engaging this group. Then check the definition in the glossary at the back of this book and in this chapter's discussion of review panels.
5. Identify a program with which you are familiar. Assuming you have been asked to conduct a summative evaluation of the program, draft an outline for the final evaluation report.
6. Identify and summarize the contents and uses of at least two checklists for guiding the reporting of evaluation findings.
7. What is visual processing theory, and what are at least three examples of its utility in preparing evaluation reports?
8. Define issues that evaluators must deftly take into account when they report to policy boards rather than to administrators, and identify at least two mistakes to be avoided when reporting to a policy board.
9. Identify potential areas of conflict in the context of leading a discussion of an evaluation report, and list at least ten of what you consider to be the most important ways an evaluator can forestall or manage conflict during the discussion.
10. Under what circumstances would an evaluator and client consider implementing an input evaluation as a follow-up study, and how would the evaluator use such a follow-up study to respond to issues involved in presenting recommendations based on the original evaluation?

Group Exercises

Exercise 1

Suppose your group is being considered to conduct an independent evaluation of a charter school that is housed in a public school district. The matrix that follows is designed to help your group explain to the prospective client—in this case the school district's superintendent—which persons should be engaged in deciding or helping decide, in general, who should be served by the evaluation and, in particular, who should be authorized to receive interim evaluation reports and who should be authorized to receive the final report.

Your group's task is to fill in each cell of the matrix by inserting one or more check marks to indicate whether someone in the role named at the top of the particular column should

have exclusive authority, shared authority, no authority, consultation opportunities, or no involvement in answering the question in the leftmost cell.

After filling in the matrix's cells with check marks, use the completed matrix to discuss the relevance to the success of the projected evaluation of clarifying, in advance, evaluator, client, and stakeholder roles in defining the evaluation's different users and differentiating their rights to receive interim reports, the final report, and evaluation services in general. In the course of the discussion, your group may find it useful to recall stipulations in Chapter 21's discussion of evaluation contracting about what parties should be authorized to negotiate and sign evaluation contracts.

	Roles in Planning an Evaluation		
Question to Be Answered in Defining an Evaluation's Appropriate Users	**Prospective Evaluator**	**Prospective Client (School District Superintendent)**	**Program Stakeholders (Charter School Teachers, Students, and Parents, Plus Other Interested Parties)**
In general, what persons should be served by the evaluation?	___ Exclusive authority to decide ___ Shared authority to decide ___ No authority to decide ___ Consultation ___ No involvement	___ Exclusive authority to decide ___ Shared authority to decide ___ No authority to decide ___ Consultation ___ No involvement	___ Exclusive authority to decide ___ Shared authority to decide ___ No authority to decide ___ Consultation ___ No involvement
What persons should be authorized to receive interim evaluation findings?	___ Exclusive authority to decide ___ Shared authority to decide ___ No authority to decide ___ Consultation ___ No involvement	___ Exclusive authority to decide ___ Shared authority to decide ___ No authority to decide ___ Consultation ___ No involvement	___ Exclusive authority to decide ___ Shared authority to decide ___ No authority to decide ___ Consultation ___ No involvement
What persons should be authorized to receive the final evaluation report?	___ Exclusive authority to decide ___ Shared authority to decide ___ No authority to decide ___ Consultation ___ No involvement	___ Exclusive authority to decide ___ Shared authority to decide ___ No authority to decide ___ Consultation ___ No involvement	___ Exclusive authority to decide ___ Shared authority to decide ___ No authority to decide ___ Consultation ___ No involvement

Exercise 2

Table 24.1 provides a format for identifying an evaluation's potential users and determining how findings will be used. Study this table, and then decide as a group the extent to which you find it useful in delineating potential users and uses of evaluation reports. Are there any other categories of audience members or potential evaluation uses you would include?

Exercise 3

Many, but not all, evaluations culminate in a final report. Discuss types of evaluation studies that logically require formal written reports, those that seldom need such formality, and those in which decisions about reporting are left open at the planning stage. In each instance, give reasons for decisions about reporting.

Exercise 4

"There are numerous ways in which evaluators are able to provide clients with follow-up support to enhance the impact of the findings contained in evaluation reports." Discuss this statement, giving special emphasis to helping client groups interpret and apply evaluation findings. Also, identify and comment on the potential utility of relevant checklists.

Exercise 5

What are the issues involved in developing recommendations based on an evaluation of a program? Under what circumstances should evaluators decline to present recommendations? Under what circumstances can evaluators develop and report defensible recommendations?

Exercise 6

Recapitulate and react to this chapter's contention that an evaluation's sounding board should be labeled a "review panel," not an "advisory panel" or "steering committee." How is this position reflected in our assessment of empowerment evaluation in Chapter 5?

Suggested Supplemental Readings

Coryn, C.L.S. (2006). A conceptual framework for making evaluation support meaningful, useful, and valuable. *Evaluation Journal of Australasia*, *6*(1), 45–51.

Evergreen, S.D.H. (2011). *Death by boredom: The role of visual processing theory in written evaluation communication*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Hendricks, M., & Handley, E. A. (1990). Improving the recommendations from evaluation studies. *Evaluation and Program Planning*, *13*, 109–117.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Joint Committee on Standards for Educational Evaluation. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2012). *Essentials of utilization-focused evaluation: A primer*. Thousand Oaks, CA: Sage.

Scriven, M. (1993). *Hard-won lessons in program evaluation*. New Directions for Program Evaluation, no. 58. San Francisco, CA: Jossey-Bass.

Torres, R. T., Preskill, H., & Piontek, M. E. (2005). *Evaluation strategies for communicating and reporting: Enhancing learning in organizations* (2nd ed.). Thousand Oaks, CA: Sage.

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

Worthen, B. R., Sanders, J. R., & Fitzpatrick, J. L. (1997). *Program evaluation: Alternative approaches and practical guidelines* (2nd ed.). New York, NY: Longman.

Yau, N. (2011). *Visualize this: The FlowingData guide to design, visualization, and statistics*. Hoboken, NJ: Wiley.

PART FIVE

METAEVALUATION AND INSTITUTIONALIZING AND MAINSTREAMING EVALUATION

We bring the book to a close in Chapters 25 and 26 by discussing the metaevaluation imperative and the processes of institutionalizing and mainstreaming evaluation. Our final messages are that evaluators (1) must, like any other group of professionals, subject their services to formative and summative evaluations for purposes of both improvement and accountability, and (2) should assist their client organizations to institutionalize an evaluation system or strengthen an existing system and to mainstream the system's application and use through all levels of the organization.

CHAPTER 25

METAEVALUATION: EVALUATING EVALUATIONS

This chapter is focused on metaevaluation, the process of evaluating evaluations.[1] Metaevaluators serve evaluators by formatively assessing their evaluation plans, procedures, reports, and impacts and also by providing summative assessments that evaluators can use to meet their own accountability requirements. Further, metaevaluators inform evaluation clients and users about an evaluation's utility, feasibility, propriety, accuracy, and accountability. Metaevaluations keyed to sound standards for evaluations are the foundation stone of sound evaluation services.

Evaluation is of fundamental importance to society, and the evaluation field has advanced substantially in regard to its professional organizations, standards, approaches, methods, levels of experience, and public service. Nonetheless, due to a host of technical, political, organizational, and psychological complications, many things can and do interfere with and threaten the success of evaluation work. A compromised evaluation might yield invalid conclusions and mislead its audience. Other evaluations operating under severe resource and time constraints or a narrow focus may generate findings that are sound but with highly restricted applicability. Clearly it is in professional and public interests that evaluators subject their evaluations to rigorous metaevaluation.

Fundamentally, a metaevaluation is an evaluation of an evaluation (Scriven, 1969b). Systematically evaluating evaluations is profoundly important, because it helps evaluators detect and address problems, ensure quality in their studies, and forthrightly reveal an evaluation's limitations. Moreover, metaevaluation reports help audiences judge an evaluation's relevance, integrity, trustworthiness, cost-effectiveness, and applicability (Stufflebeam, 2001b).

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **Metaevaluation's rationale and definition**
- **The case for grounding metaevaluations in professional standards for evaluations**
- **Caveats in choosing and applying published standards and guiding principles for evaluations**
- **A conceptualization of formative and summative metaevaluation**
- **The need for both internal and external metaevaluations**
- **Qualifications needed to undertake metaevaluations**
- **Evaluator and client responsibilities in conducting metaevaluations**
- **The contrast between metaevaluation and meta-analysis**
- **Metaevaluation tasks and associated checklists**
- **The special case of comparative metaevaluations**

We stress, at this early point in the chapter, that fully functional and credible metaevaluation services should include both an internal self-assessment of an evaluation by the evaluator and an external assessment by an independent metaevaluator.

We posit in this chapter that evaluators should conduct metaevaluations to guide evaluation planning and implementation and to report an evaluation's strengths and weaknesses, and that an evaluation's client should procure an independent metaevaluation to assess the completed evaluation's soundness. Systematic evaluation has become increasingly important in helping entities in government and private sectors diagnose and address such problems as failing schools; bankrupted cities; unsafe transportation services; breakdown of the family; environmental pollution; illegal immigration and border control; drug abuse; gang violence; lack of gun control; soaring health care costs; lack of health insurance for the poor; unemployment; government fraud, waste, and abuse; deterioration of bridges and other infrastructure; and terrorism. Sound evaluations are needed not only to assess the costs and outcomes of improvement and problem-solving efforts but also to help target, plan, and guide such efforts. Of course, the needed evaluations must be relevant and technically sound, and this is where formative metaevaluations play a crucial role. Typically, such formative metaevaluations will be performed by the evaluator who is conducting the evaluation, but in some cases that evaluator or preferably the evaluation's sponsor may also engage an independent metaevaluator to conduct formative as well as summative metaevaluations of the evaluation.

Rationale for Metaevaluation

As with other professional enterprises, an evaluation can be excellent, poor, or mediocre. Many things can and do go wrong in evaluations. They might be flawed due to, for example, inadequate focus, inappropriate criteria, poor design, unreliable measuring devices, a deficient contract, insufficient resources, an unrealistic time limit, incompetent evaluation personnel, participants with serious conflicts of interest, poor oversight and coordination, uncooperative program personnel, subterfuge or even sabotage by program stakeholders, invalid information, recording or analysis errors, excessive costs, late reports, ambiguous reports, biased findings, unsupported conclusions, unwarranted recommendations, or corrupt or misguided uses of findings. Such problems may occur across the full range of disciplines and service areas.

If such problems are not detected and addressed in the evaluation process, and if the evaluation survives to a conclusion, an evaluator may present erroneous findings; deliver services that are overly expensive, ineffective, or unfair; or become complicit in unethical or misguided uses of findings. Further, the evaluation will probably impair the evaluator's credibility. If flawed reports are issued without being exposed as such by sound metaevaluations, evaluation audiences may make bad decisions based on the erroneous findings. As Scriven (1994d) reported, even the highly respected and widely used *Consumer Reports* magazine should be independently evaluated to help readers see the limitations as well as the strengths of the many product evaluations published therein. Further, invalid personnel evaluations can have unfair or inappropriate consequences for employees or, conversely, for an institution and its clients (Brannick & Levine, 2002; Brannick, Levine, & Morgeson, 2007). Similarly, evaluations

that accredit unworthy programs or institutions are a disservice to potential students or other constituents. In addition, professional standards and principles for evaluations will be little more than rhetoric if they are not applied in the process of judging and improving evaluation services.

Metaevaluations are in the best interest of the public, professionals, and institutions to ensure that evaluations provide sound findings and conclusions; that evaluation practices continue to improve; and that institutions administer efficient, effective, ethical evaluation systems. Evaluation has developed as a vital area of professional service as societal groups have increasingly engaged evaluators to investigate and pass judgment on many and varied consumer programs, projects, products, policies, organizations, and services as well as to conduct needs assessments, diagnostic investigations, and implementation studies to help groups focus, plan, and carry out needed interventions.

Metaevaluation is a professional obligation of evaluators. Achieving and sustaining the status of a professional requires subjecting one's work to evaluation and using the findings to ensure sound services and to strengthen services over time. This dictum pertains as much to evaluators as it does to accountants, architects, engineers, lawyers, U.S. Postal Service officials, government program administrators, hotel and restaurant managers, judges, public safety officials, school and university administrators, air traffic controllers, construction contractors, electricians, plumbers, emergency care workers, religious clerics, nurses, physicians, dentists, pharmacists, psychologists, teachers, national defense officers, and other service providers. It means that evaluators should ensure that their evaluations are themselves evaluated. Moreover, metaevaluations are needed in all types of evaluation, including evaluations of programs, projects, products, services, budgets, expense reports, equipment, systems, organizations, theories, models, policies, research designs, artistic works, conferences, students, and personnel and across the full range of professions and service fields.

Apart from needing metaevaluations to ensure the quality of their evaluations, evaluators, as professionals, should use metaevaluations to provide direction for improving their developing evaluation approaches and tools and to earn and maintain credibility for their services among client groups and other evaluators. Consumers need metaevaluation reports to help decide whether to accept and act on evaluative conclusions about products, programs, services, and other evaluands they are using or considering for use. They need metaevaluation reports to identify sound evaluations, use those evaluations' findings with confidence, and know which corrupt or faulty reports to disregard. Basically, cogent, defensible metaevaluations help users avoid accepting invalid evaluative conclusions, instead enabling them to make measured, wise use of sound evaluative information. Those who may not aspire to be a professional evaluator but who house and oversee an evaluation system need metaevaluations to help ensure that their institution's evaluation services are relevant, ethical, technically sound, practical, usable, timely, efficient, and worth the investment. Also, the subjects of evaluations—including program staff members, contractors, various other professionals, and students—have the right to expect that the system used to evaluate their qualifications, experience, competence, and performance meets appropriate standards for sound personnel evaluations or student evaluations. Clearly, metaevaluations are in the best interest of a wide range of parties who may be served or affected by evaluations (Scriven, 2009b).

Evaluator and Client Responsibilities in Regard to Metaevaluation

We wish to state emphatically that clients and funders, as well as evaluators, bear responsibility for obtaining and using sound metaevaluations. Clearly, evaluators should key their evaluations to standards for sound evaluation and should conduct both formative and summative metaevaluations, the former to guide the evaluation process and the latter to attest to the soundness of the evaluation process and findings and to acknowledge limitations and deficiencies. However, the evaluator has a conflict of interest in assessing her or his own evaluation; therefore, typically the evaluation's client or funder should commission, fund, use, and release to right-to-know audiences an independent summative assessment of the evaluation. We cannot stress this point too strongly, because in far too many cases an independent metaevaluation has not been conducted or the evaluator has chosen a "friendly critic" to conduct the metaevaluation—one who may be expected to produce an overly favorable report. Moreover, we do not think a funder or client is justified in failing to commission an external metaevaluation due to the issue of cost; in our experience, a sound metaevaluation constitutes only a fraction of a study's overall cost and can deliver findings for the funder or client and other right-to-know audiences whose benefits more than compensate for the investment in the metaevaluation.

The rub is how to convince clients and funders that, beyond funding a primary evaluation, they should also commission and fund an independent metaevaluation of the subject evaluation. We lay the responsibility for making this case to the client or funder on the primary evaluator. That is, in contracting for an evaluation, the evaluator should strongly advise the client or funder to commission and fund an independent metaevaluation and should stress to the client or funder that the metaevaluation should be keyed to professional standards for sound evaluations, such as those associated with the evaluation's utility, feasibility, propriety, accuracy, and accountability (Joint Committee on Standards for Educational Evaluation, 2001; also see Cooksy & Caracelli, 2005, 2009; Scriven, 2009b).

Formative and Summative Metaevaluations

Proactive metaevaluations are needed to help evaluators focus, design, budget, contract, and carry out sound evaluations. Retrospective metaevaluations are required to help audiences judge completed evaluations. In the evaluation literature, these two kinds of metaevaluation are referred to as formative metaevaluation and summative metaevaluation, respectively (Stufflebeam, 2000b). As already noted, we see a need for both internal and external formative and summative metaevaluations. This notion is summarized in Table 25.1.

A Conceptual and Operational Definition of Metaevaluation

The practice of evaluating evaluations has a long history in fields related to evaluation, although the term *metaevaluation* has been applied to this area only since 1969. Over the years, evaluation reports on such national issues as racial segregation, surreptitious monitoring of the long-term effects of syphilis on African American military personnel, ensuring gender equality, abuses of achievement testing, benefits of preschool education, busing schoolchildren to achieve school integration, the quality of schools, the link between smoking and lung

Table 25.1 Framework for Internal and External Formative and Summative Metaevaluations

	Formative Metaevaluation	Summative Metaevaluation
Internal	The evaluator systematically assesses and takes needed steps to clarify or strengthen the evaluation's definition of the target audience, questions, plans, budget, negotiated agreement, tools, personnel, data collection, data analysis, draft reports, and so forth.	The evaluator appends to the final evaluation report an attestation to the extent to which the evaluation met appropriate professional standards (for example, those associated with the evaluation's utility, feasibility, propriety, accuracy, and accountability).
External	The independent metaevaluator, chosen by the evaluation's client or funder, monitors the unfolding evaluation and, as appropriate, provides the evaluator with critical feedback on the evaluation's strengths and weaknesses.	The independent metaevaluator, chosen by the evaluation's client or funder, compiles and delivers a summative metaevaluation report on the completed evaluation, assessing its utility, feasibility, propriety, accuracy, and accountability. Said report should be made available to all intended users of the subject evaluation.

cancer, the link between asbestos and lung cancer, the safety of mining practices, the safety of thalidomide, oil drilling and fracking to achieve energy independence, global warming, and the link between paper mills and water quality in rivers and the Great Lakes have spawned great societal debates, especially concerning the need for new laws and regulations. In turn, these debates have led to evaluations of the subject evaluation reports. Clearly such metaevaluations have been crucially important in helping public officials and the public adjudicate the validity of various evaluations' reported conclusions and recommendations and sometimes in helping government officials enact responsive legislation or take other corrective actions.

Scriven (1969b) introduced the term *metaevaluation* in *Educational Products Report*. He employed this label to refer to his evaluation of a plan for evaluating educational products. Scriven essentially defined a metaevaluation as any evaluation of an evaluation, evaluation system, or evaluation device. He argued that issuance of inaccurate or biased reports can seriously mislead consumers into purchasing unworthy or inferior educational products, which they might then use to the detriment of children and youth. Thus, he stressed, the evaluations of such products must themselves be evaluated, and such metaevaluations are critically important to the welfare of consumers. Although Scriven initially focused the term *metaevaluation* on the narrow sphere of evaluations of educational products, it is clear that the underlying concept is applicable to the widest possible range of evaluations. Every evaluation study should be sound, and its soundness should be ensured and enhanced through formative metaevaluation and verified or discredited through one or more defensible summative metaevaluations. As we stressed earlier, there are important reasons for clients and funders to secure and use independent metaevaluations and not to rely only on the evaluator's attestation to the quality of her or his final evaluation report.

An Operational Definition

Operationally, we define metaevaluation as the process of delineating, obtaining, and applying descriptive and judgmental information about an evaluation's utility, feasibility, propriety, accuracy, and accountability for the purposes of guiding the evaluation and reporting its strengths and weaknesses.

This definition is designed particularly to serve the metaevaluation needs of evaluators, clients, and funders who choose to adhere to the standards issued by the Joint Committee (1981, 1988, 1994, 2003, 2011). The definition is also consistent with metaevaluation applications of the 2004 American Evaluation Association (AEA) guiding principles for evaluators and the 2007 U.S. Government Accountability Office (GAO) government auditing standards. (See Chapter 3 for detailed information on alternative sets of standards that are applicable to metaevaluations.)

It is instructive to consider the main elements of the definition of metaevaluation, especially from the perspective of an independent metaevaluator. The process elements of this definition include both group process and more discrete technical tasks. The group process tasks of delineating and applying denote a metaevaluator's interactions with an evaluation's evaluator, the client, and other stakeholders of the evaluation being assessed. In planning the metaevaluation, the metaevaluator identifies and communicates and negotiates with the client and stakeholders, as appropriate, to reach mutual understandings on the important metaevaluation questions, how they are to be addressed, how and when the findings are to be reported, areas of responsibility and authority in regard to different aspects of the work, resources for the engagement, and the metaevaluation standards. In the metaevaluation's concluding stages, the metaevaluator meets or otherwise communicates with the client and other stakeholders to help them understand, correctly interpret, and apply the metaevaluation findings. Typically the metaevaluation findings are also conveyed to the evaluator whose work was assessed. In presenting the metaevaluation findings, the metaevaluator should help audience members draw justified conclusions and should caution them against overgeneralizing or otherwise incorrectly interpreting or inappropriately using findings.

The obtaining elements in the definition of metaevaluation are the technical tasks required to collect, analyze, and synthesize the information needed to judge the target evaluation. Especially involved are the collection and assessment of an evaluation's plans, budget, contract, instruments, data, implementation records, interim and final reports, and expense reports; evaluator credentials; and evidence of uses of findings. In addition, a metaevaluator may interview, survey, or otherwise collect information and perspectives from persons involved in or affected by the evaluation process. As the definition notes, a metaevaluation should be informed by both descriptive and judgmental information. Descriptive information may include statistics and other quantitative information as well as qualitative information as might be gathered from interviews and content analysis. Pertinent judgmental information includes judgments by stakeholders and other interested parties of any and all aspects of the subject evaluation and might be obtained through focus groups, interviews, hearings, newspaper editorials, and the like. It is noteworthy that many metaevaluators fail to stay engaged long enough to identify and assess an evaluation's impacts (or lack of impacts) pertaining to such future decisions and actions as increasing or decreasing funding, changing procedures or timelines, introducing efficiency measures, reformulating policies, training personnel, disseminating lessons learned, and replicating the program elsewhere (Cooksy & Caracelli, 2005).

Our definition's bases for judging evaluations are the standards for program, personnel, and student evaluations developed by the Joint Committee (2003, 2009, 2011). In addition, we

advise evaluators to use other sets of professionally developed standards that are appropriate in particular evaluations. In particular, these could include AEA's *Guiding Principles for Evaluators* (2004); GAO's *Government Auditing Standards* (2007); and the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education's *Standards for Educational and Psychological Testing* (1999) for evaluating educational achievement tests and psychological instruments. In some situations, it will be instructive to employ more than one set of standards for assessing an evaluation. Notably, J. R. Sanders (1995) found that the 1994 Joint Committee program evaluation standards and the AEA guiding principles for evaluators (Shadish, Newman, Scheirer, & Wye, 1995b) are compatible and complementary. Grasso (1999) merged and jointly applied both sets of requirements. In their analysis of eighteen metaevaluations, Cooksy and Caracelli (2009) found that the employment of criteria for metaevaluations ranged from the use of emergent criteria in a narrative review of information about an evaluation to the structured application of the 1994 Joint Committee program evaluation standards using a checklist.

Qualifications to Undertake Metaevaluations

Professionally recognized standards provide the foundation for sound metaevaluations. However, the irksome question lingers: Who is qualified to carry out such an important task? What metaevaluator qualifications should clients or funders bear in mind as they search for and choose, commission, and fund someone to conduct an independent metaevaluation? We believe that the following qualifications are essential:

- *A working knowledge of alternative sets of professional standards for evaluations, together with a demonstrated ability to choose, adapt as appropriate, and apply standards that fit particular evaluation assignments.* Without such sure reference points, a metaevaluator's decisions will appear arbitrary and lack credibility. Moreover, a metaevaluator and audience will lack a common, valid basis for adjudicating findings and conclusions. (The next subsection underlines serious potential problems resulting from the inappropriate application of standards, and also emphasizes that adapting standards may be acceptable if circumstances so warrant.)
- *A combination of methodological expertise and comprehension of the subject program or other evaluand.* A metaevaluator must be qualified to investigate and judge the technical aspects of the subject evaluation, and also should possess or develop substantial knowledge of the substantive domain of the subject program or other evaluand. Of course, a metaevaluator must be capable of judging an evaluation's quantitative and qualitative methods. Without adequate grounding in an evaluation's substantive domain, a metaevaluator is apt to produce findings that are superficial, invalid, or viewed with suspicion. Because metaevaluators perform their services across a wide spectrum of evaluations and substantive areas, however, they cannot be expected to enter every metaevaluation assignment with the needed level of content expertise. More often than not, a metaevaluator will begin a metaevaluation assignment with limited knowledge of

the evaluand on which the target evaluation is focused. Metaevaluators must therefore excel in making quick, extensive studies of pertinent subject matter domains. Moreover, in certain cases a client or funder will need to engage a metaevaluation team whose members possess the needed combination of methodological and content area expertise.

- *Sufficient experience and competence to meet clients' metaevaluation needs*. Prospective metaevaluators should provide evidence of competent performance as metaevaluators or present credentials to show that they are well trained in the concepts, standards, and methods of metaevaluation. In either case, they should present documentation affirming that they possess the knowledge and ability needed to meet clients' needs. Skills and experience in undertaking evaluations confidently and effectively are essential, basic requirements, and it is highly desirable that the prospective metaevaluator have previously delivered effective, independent metaevaluations.
- *Honesty, integrity, and respect for individuals and society generally*. These qualities are essential for all evaluators and have special significance for metaevaluators. Most metaevaluations culminate in a final judgment of an evaluation. At stake are uses of a subject evaluation and also the reputations of the evaluator of the program or other evaluand and of program leaders, managers, and staff. Especially affected are decisions based on the subject evaluation, which may be overturned or sustained in response to metaevaluation findings. The consequences of a metaevaluation's misrepresentation of the subject evaluation for whatever reason may be dire. The metaevaluator's craft involves validly substantiating or discrediting an evaluation's findings and conclusions, while respecting and guarding against bringing undue harm to the dignity and self-esteem of involved parties. Complex issues often are involved because of the diversity of stakeholders and their interests. This fact underscores the imperative that metaevaluators be sensitive, ethical, and competent.
- *Skills in negotiating formal metaevaluation contracts*. Such contracts should include clarification of a metaevaluation's client, appropriate audiences, the budget, standards and methodologies to be employed, guarantees of access to needed information, and responsibility for and authority over editing and distributing reports. Other contractual factors may include follow-up services to help ensure that the metaevaluation has an impact. (See Chapter 21 for details on contracting.)
- *An ability to effectively communicate and collaborate with other parties to a metaevaluation*. Typically a metaevaluator or metaevaluation team will have involvements with a wide range of persons and groups: a program's evaluator, director, staff, and beneficiaries, among others. Metaevaluators require skills to communicate forthrightly and diplomatically with this full range of stakeholders. Especially, they should be skilled in putting at ease and having substantive exchanges with persons who might well be threatened by the metaevaluation or intimidated by an external metaevaluator. Metaevaluators should show respect for the full range of parties to a metaevaluation and deal evenhandedly with all of them. Moreover, they should be able to stimulate interest in the metaevaluation and generate a spirit of cooperation among the entire group toward obtaining valid, useful findings. They should be skilled in developing rapport with contributors to the metaevaluation and in asking incisive

initial and follow-up questions. Metaevaluators must be excellent listeners. Moreover, they should take care to ensure that those stakeholders who contribute information for or otherwise assist with the metaevaluation are compensated by receiving or having access to the metaevaluation's findings.

Caveats in Applying Standards and Principles

Some caveats are in order concerning the use of standards and guiding principles. The standards released by the Joint Committee (2003, 2009, 2011) are focused on evaluations of educational and training programs, education personnel, and students, and are not designed for use in evaluating programs and other objects outside the fields of education and training. Moreover, the Joint Committee expressly developed its standards for use in the United States and Canada and pointedly cautioned against uncritical use of these standards outside the North American context. In this regard, it should be remembered that the Joint Committee defined a standard as a principle commonly agreed to by the parties whose work will be evaluated against the standard. The committee argued that evaluators in other countries should carefully consider what standards are acceptable and functional within their culture. There are definite problems in transferring North American standards on human rights, freedom of information, the right to privacy, and other matters covered by the Joint Committee's standards to cultures outside the United States and Canada (Beywl, 2000; Jang, 2000; N. L. Smith, Chircop, & Mukherjee, 2000; Taut, 2000; Widmer, Landert, & Bacmann, 2000). Nevertheless, evaluators in a number of countries have adapted the Joint Committee's standards for use in guiding and judging their evaluations and secured agreements among their colleagues to apply the standards. Also, evaluators often have adapted the Joint Committee's standards and used them to conduct evaluations outside the fields of education and training. When evaluators and their client groups agree that the Joint Committee's standards are applicable to their noneducational evaluations, we see no problem with pertinent adaptations and applications.

The 2004 AEA guiding principles for evaluators also have an American orientation. It is noteworthy, however, that AEA's membership includes evaluators from many countries. AEA seems to espouse the reasonable position that its members and other evaluators, wherever they may be conducting evaluations, should adhere at least to the AEA guiding principles if they intend to claim consistency between their evaluations and what AEA recommends for conducting sound evaluations. It is noteworthy that the Joint Committee's standards provide extensive detail beyond what is included in the AEA guiding principles, but that the guiding principles are designed for use across a wide array of disciplines and service areas. It often makes sense to apply both the Joint Committee standards and the AEA guiding principles in metaevaluations, thus taking advantage of their complementarities.

GAO (2007) has shared its government auditing standards with government accounting organizations and accountants throughout the world, for whatever considered use they might make of these U.S. standards. The GAO standards are referenced and used, at least as a model, in a wide range of countries. We perceive that many auditors throughout the world view the GAO standards as the best available model for setting and applying standards of financial accounting in government programs.

The point of this discussion of caveats is that standards used in given metaevaluations should have been validated for such use. Clearly, metaevaluators can choose from alternative sets of standards for sound evaluations. In Chapter 3 we presented the standards that we recommend for consideration; but these standards should be used in accordance with their stated purposes and spheres of applicability. Although these standards provide excellent examples of essential aspects of standards, they are not intended for universal applicability. Evaluators outside North America should carefully determine which standards are professionally and politically acceptable in their context. These might or might not reflect the contents of the standards that have been developed and adopted for use in the United States and Canada.

The final part of this chapter's operational definition of metaevaluation highlights its basic purposes as guiding the evaluation and reporting its strengths and weaknesses. Like any other kind of evaluation, a metaevaluation may have a formative role in helping an evaluation succeed and a summative role in helping interested parties judge the evaluation's merit and worth.

Metaevaluation in Relation to Meta-Analysis

As part of our conceptualization of metaevaluation, it is instructive to contrast metaevaluation and meta-analysis. Although these terms refer to quite different concepts, they are often inappropriately equated. A metaevaluation assesses the merit and worth of a given evaluation, evaluation system, or evaluation device. A meta-analysis is a form of quantitative synthesis of studies that address a common research question (see Chapter 6). In program evaluation research contexts, a meta-analysis usually involves a contrast between treatment and control or between treatment A and treatment B. Conducting research across a selected set of similar studies, an investigator calculates and examines the positive or negative direction, magnitude, and statistical significance of effect sizes. The objective is to determine the pervasive, overall effect of a class of treatments.

Although metaevaluation and meta-analysis are different activities, metaevaluations have applications in meta-analysis studies (also see T. D. Cook & Gruder, 1978). Metaevaluations are used first to evaluate and determine which candidate comparative studies qualify for inclusion in a defensible meta-analysis database. Also, a metaevaluation can and should be conducted to assess the merit and worth of a completed meta-analysis. The meta-analysis technique is rarely applicable in a metaevaluation, because most evaluations do not involve multiple comparative studies in a particular program area.

An Instructive Metaevaluation Case

The following example describes a metaevaluation that adhered quite closely to the preceding conceptualization of metaevaluation. Our presentation of this metaevaluation is intended to help readers identify the main tasks that are present in most metaevaluations.

This illustration is of a metaevaluation of the teacher evaluation system previously employed by Teach for America (TFA), an organization that recruits, trains, and certifies

graduates of various baccalaureate programs for service as teachers in inner-city schools. In our illustration, the teacher trainees had four-year degrees grounded in an arts and sciences discipline, but most had no university-based teacher education. TFA's role was to recruit able college graduates desiring to serve inner-city students; provide them with a year of on-the-job, supervised teacher training in inner-city schools; rigorously evaluate their performance and potential during and immediately following this probationary period; and subsequently recommend only satisfactory performers for certification as effective teachers.

The metaevaluation function was important to TFA's financial sponsors, administrators, certifying bodies, constituent school districts, and teacher trainees. TFA's leaders needed to demonstrate the program's quality, integrity, and ability to produce and recommend only competent teachers, because the program was an innovative, radical alternative to traditional programs for educating and certifying teachers. Those other programs involved teacher trainees in a four-year or five-year, on-campus college or university baccalaureate program. Some members of the teacher education establishment charged that TFA's program was inferior to traditional college and university teacher education programs, because the teacher trainees had not received sustained instruction from teacher educators in school functioning, teaching techniques, classroom management, assessment techniques, and characteristics of children and youth. Such criticisms are ironic because traditional teacher education programs have been severely criticized for failing to produce sufficient quantities of keenly intelligent, effective instructors with a substantial grounding in relevant disciplines. Clearly the traditional programs had not supplied enough certified teachers to meet inner-city school districts' needs, as evidenced by the large numbers of vacant teaching positions; employment of many teachers with emergency, provisional certificates; assignment of teachers to teach subject matter outside their college major; and the heavy use of substitute teachers.

Urban school districts have had great difficulties in hiring and retaining competent teachers to fill needs in the various subject matter and specialty areas and across grade levels. With the advent of TFA, many inner-city schools were understandably interested in the possibility of hiring TFA graduates to meet their staffing needs, especially because these graduates had majors in such areas as mathematics, physics, biology, chemistry, history, English, modern languages, and geography. The graduates' desire and availability to teach in inner-city schools would be a plus if they also possessed skills in regard to classroom management, instructional planning, effective communication, motivating students, using technology, achievement testing, teamwork, and securing parental involvement. Urban school districts potentially were an important market for TFA.

Understandably, the school districts and their respective school boards required solid assurances that TFA's graduates were well trained and appropriately equipped to deliver excellent teaching services to inner-city students. State bodies that certify teachers also needed evidence that TFA's teacher candidates had developed the needed teaching competencies. The teacher candidates themselves needed assurances that they would be functioning in a profession to which they were well suited and for which they were appropriately prepared, and that they would be welcomed and respected. They also deserved to be credentialed or

screened out based on fair, valid, impartial assessments. If TFA could expect state governments to approve, trainees to enroll, and schools to employ the graduates, it needed to achieve and maintain credibility based on the soundness of its bold alternative to traditional teacher preparation and certification programs. Fundamentally it needed to do an excellent job of preparing the teacher candidates and also conduct and report on a rigorous, credible evaluation of the postpreparation qualifications of each TFA graduate. TFA's evaluation of the teacher trainees was a crucial task in awarding certification and getting only qualified teachers into the schools. Its system for evaluating the probationary teachers was labeled the performance assessment system (PAS).

In 1995 TFA commissioned a metaevaluation to help ensure that PAS was providing a credible, technically sound basis for assessing the competence of TFA's teacher candidates. The metaevaluation also was deemed important to convince interested parties that PAS had been subjected to an independent metaevaluation. TFA engaged William Mehrens from Michigan State University, Jason Millman from Cornell University, and Daniel Stufflebeam from Western Michigan University to serve as the metaevaluation team. All three were extensively published experts in evaluation theory and methodology and collectively possessed specialized expertise in state teacher certification systems, measurement and statistics, evaluation standards, and metaevaluation procedures. TFA commissioned this team to determine whether PAS, in design and execution, fairly, reliably, and accurately evaluated beginning teachers.

The five main components of PAS were teacher-compiled portfolios, portfolio assessors, a system of training and calibrating the assessors, systematic examination and assessment of portfolios, and certification recommendations derived from the assessments. The evidence in each teacher's portfolio included teaching plans, videotaped teaching, the teacher's assessment devices, student work, and an analysis of the students' academic growth. The portfolio also included survey results from the teacher's principal, other supervisors, teacher colleagues, parents, and students.

Two specially trained assessors independently evaluated each portfolio according to preestablished rubrics and produced subscores for the specified certification criteria. A third assessor resolved any unacceptable discrepancies between the first two sets of ratings.

The metaevaluators assessed whether each of the following was sound: the performance assessment design and criteria, assessors' selection and training, implementation of the portfolio review process, assessments of teachers' impacts on student learning, quantitative analysis of the assessors' ratings of probationary teachers, legal defensibility of PAS, and plans for PAS's wider use. The metaevaluators also assessed PAS against the requirements of the 1988 Joint Committee personnel evaluation standards to reach judgments about PAS's utility, feasibility, propriety, and accuracy.

The metaevaluators first obtained from TFA and studied documents related to the selected metaevaluation questions and the twenty-one 1988 Joint Committee personnel evaluation standards. Among others, these documents included the teacher trainees' academic records; the credentials of the assessors assigned to evaluate the evidence on each probationary teacher; and the TFA plan and associated recruitment, training, and assessment criteria and

devices. The metaevaluators observed and prepared field notes on the training of assessors and examined a sample of the beginning teachers' portfolios. Subsequently they observed the assessors' actual assessments of the teachers' materials and analyzed the ratings and resulting certification recommendations, especially for reliability of subscores and agreements on final recommendations. Throughout the process, the metaevaluators conducted both telephone and face-to-face interviews with a range of participants. Examination of the obtained evidence was used to judge whether TFA's PAS met, partially met, or failed to meet each of the 1988 Joint Committee standards. The metaevaluators also referenced pertinent policies, statutes, and laws to assess the legal viability of TFA's assessment structure and process. Finally, they produced an executive summary, a full-length report, and a technical appendix for the completed metaevaluation. In accordance with the metaevaluation contract, these reports were delivered to TFA for its discretionary use. The metaevaluation contract included no provision for follow-up work by the metaevaluators—a limitation, if not a deficiency, of the metaevaluation.

The basic findings were that TFA's evaluation team performed creditably and legally in conducting and reporting summative evaluations of probationary teachers. Also, TFA was judged to have performed professionally in informing its state department and school district clients about the evaluation findings. The metaevaluation identified areas where TFA needed to improve PAS, especially in providing less hurried training for the assessors, strengthening the assessments of the teacher trainees' impacts on student learning, and better matching assessors and trainees on content areas and grade levels taught.

Metaevaluation Tasks

The preceding TFA example points to eleven main tasks that should be considered in planning a metaevaluation process. These tasks are summarized in Table 25.2 in terms of the eleven task areas and the definition of each one's central task. Basically, these are generic tasks. Because metaevaluation is only a special type of evaluation, readers should not be surprised that the identified tasks apply to evaluations in general and not only to metaevaluations. Readers should keep in mind that these tasks were derived from the PAS metaevaluation case; not all the tasks would necessarily apply in all formative and summative metaevaluations, and additional tasks sometimes are needed. The following task areas and associated tasks, then, are suggested mainly as a heuristic for use in planning metaevaluations and selecting appropriate methods.

Task Area 1: Staffing

TFA selected Millman, Mehrens, and Stufflebeam to address three specialized areas, in addition to the broader area of teacher evaluation systems, in which they had previous experience. Millman, a past president of the National Council on Measurement in Education, focused especially on technical measurement questions, a topic on which he had published much and was eminently qualified. Mehrens examined PAS's legal viability, reflecting his extensive experience in assessing the legal viability of state teacher certification systems. Stufflebeam assessed PAS against the twenty-one 1988 Joint Committee personnel evaluation standards,

Table 25.2 Generic Metaevaluation Tasks

Task Area	Task
1. Staffing	Staff the metaevaluation with one or more qualified metaevaluators.
2. Stakeholder engagement	Identify and arrange to interact with the metaevaluation's stakeholders.
3. Standards	Agree on standards, principles, or criteria to judge the evaluation system or particular evaluation.
4. Questions	Define the metaevaluation questions.
5. Formal agreements	Issue a memo of understanding or negotiate a formal metaevaluation contract.
6. Existing information	Collect and determine the adequacy of pertinent, available information.
7. New information	Collect new information as needed.
8. Analysis and synthesis	Analyze and synthesize the obtained information.
9. Reaching conclusions	Judge the evaluation or evaluation system in terms of its adherence to appropriate standards, principles, or criteria.
10. Reporting	Convey the findings through reports, correspondence, oral presentations, workshops, and other means.
11. Follow-up	As appropriate and feasible, help the client and other stakeholders interpret and apply the findings.

whose development he had led. These considerations aside, clearly TFA did not select these team members for their gender or racial diversity, although these are often relevant considerations. Presumably this team was selected to provide expertise that the client group lacked and to provide an independent, credible perspective on TFA's evaluation process. However, clients will not always be able to or need to engage an independent metaevaluator. In some resource-poor evaluations and especially in formative evaluations, evaluators appropriately might do much or all of the formative metaevaluation themselves. Such self-metaevaluation practice is better than conducting no metaevaluation at all, provided that the evaluator systematically addresses and adheres to appropriate professional standards for evaluations.

Task Area 2: Stakeholder Engagement

In contemplating conducting a metaevaluation, the metaevaluator should clearly identify the client and other appropriate audiences for the metaevaluation reports. In the TFA case, the client group included TFA's leaders and staff. The metaevaluation also had many additional stakeholders, among them TFA's teacher trainees and the participating state education departments and school districts. Other audiences were teachers, school board members, administrators, and students and their parents in the involved school districts.

Task Area 3: Standards

A fundamentally important task is to reach agreement on the standards, principles, or criteria that will form the basis for judging the subject evaluation or evaluation system. As noted earlier, the metaevaluators and the TFA leaders agreed early in the metaevaluation planning process that PAS would be judged against the twenty-one 1988 Joint Committee personnel evaluation standards.

Task Area 4: Questions

Basically, the metaevaluators and client group in the TFA case agreed that the metaevaluation should address questions concerning the soundness of all key PAS components, especially the following:

- The performance assessment design and criteria
- Assessors' selection and training
- Implementation of the portfolio review process
- Assessments of teachers' impacts on student learning
- Quantitative analysis of the assessors' ratings of probationary teachers
- Legal defensibility of PAS
- Plans for PAS's wider use

A pervasive metaevaluation question, associated with those just listed, asked whether the individual components and PAS overall were meeting the 1988 Joint Committee personnel evaluation standards' requirements for utility, feasibility, propriety, and accuracy.

Task Area 5: Formal Agreements

Another early task in the metaevaluation process is to clarify in writing the agreements needed to guide the metaevaluation and often to negotiate a formal metaevaluation contract. Among the important agreements reached in the TFA case were the standards to use in judging the evaluation system, definition of the metaevaluation issues and questions, guaranteed access to the needed information, required substance and timing of reports, designated responsibility and authority to edit and release the metaevaluation reports, and provision of the required resources. Formal contracts are not always required, especially in small, formative, internally conducted metaevaluations. In general, though, it is wise and can prove important to clarify and record as much as feasible the basic agreements that will guide and govern the metaevaluation, including provisions for subsequently modifying the agreements by mutual consent as needed. In the contracting process, the metaevaluator and client should carefully consider whether the metaevaluator should help the metaevaluation audience interpret and apply findings following delivery of the final report. If an agreement is reached to obtain such follow-up metaevaluation work, the metaevaluation budget should include funding for this activity. Increasingly, we see follow-up metaevaluation services as crucially important to help foster effective use of metaevaluation findings and actually to enhance and document the cost-effectiveness of the metaevaluation work.

Task Area 6: Existing Information

The next task in the TFA metaevaluation was to compile and review the available, relevant information. Often a metaevaluator can collect such information at a program's central office

or even have such information delivered by mail or e-mail. The initial information collection process typically culminates in a desk review of relevant documents and filed information. In the evaluation of PAS, the metaevaluators obtained and reviewed a wide range of documents pertaining to all parts of PAS: the performance assessment design and criteria, assessors' credentials, assessor training materials, trainees' portfolios, ratings of trainees, and so on.

Task Area 7: New Information

Following review of extant information, a metaevaluator often must collect additionally needed information on-site, especially information that can be obtained only where program activities are under way. For example, the TFA metaevaluation included telephone and on-site interviews, observations of assessor training sessions, and study—in a secure setting—of portfolios collected from TFA trainees. To reach valid conclusions, metaevaluators need access to all the relevant, available information and authorization to collect any additionally needed information. Basically, they should obtain the full range of information required to address the metaevaluation questions and apply all the applicable standards and principles.

Task Area 8: Analysis and Synthesis

After compiling the needed information, a metaevaluator should analyze and synthesize the information. In the TFA case, the metaevaluators began the analysis work by summarizing each metaevaluation method's findings for each of the key metaevaluation questions and for each metaevaluation standard. Subsequently, the metaevaluation team looked across the findings for the different methods and summarized their agreements and disagreements related to answering each metaevaluation question and determining whether in the aggregate each metaevaluation standard was met, partially met, or not met.

Task Area 9: Reaching Conclusions

Based on the analysis and synthesis of findings, the metaevaluators in the TFA case then deliberated to converge on their compressed set of basic conclusions. They judged each of PAS's components and also reached judgments of PAS overall in relation to each of the employed 1988 Joint Committee personnel evaluation standards. As noted earlier, the team reached chiefly positive bottom-line judgments of PAS but also identified areas for improvement.

Task Area 10: Reporting

Reporting effectively to the client group is crucially important to secure appropriate metaevaluation impacts. Such reporting should include an executive summary suitable for wide distribution, a detailed report keyed to answering the metaevaluation questions and judging the evaluation or evaluation system against the adopted metaevaluation standards, and a technical appendix or technical report describing the tools and data that led to the metaevaluation's conclusions. Moreover, reporting should go beyond delivery of printed reports to include such

interactive activities as oral presentations and follow-up discussions, workshops to go over the findings, focus groups, and webinars.

In addressing the reporting task, the TFA metaevaluators divided up the writing assignments, produced draft sections for the final report, and reviewed and discussed the report's draft components. One team member then compiled the entire report and submitted the semifinal draft to the client for review. After considering critiques from the client and other stakeholders, the metaevaluators finalized their report and transmitted it to the client. In their printed report, the TFA metaevaluators presented tables showing both quantitative and qualitative analyses, followed by judgments of TFA's adherence to each of the twenty-one 1988 Joint Committee personnel evaluation standards. Their report generally endorsed the merit and worth of PAS, as assessed against the standards, and also pointed to specific areas requiring improvement.

As with other tasks, how extensive the reporting work needs to be depends on the metaevaluation context. For example, if the metaevaluation is sensitive, large scale, and summative, formal written reports of findings will be required along with a supporting technical appendix or a technical report. In more formatively oriented metaevaluations, however, findings may appropriately be conveyed through e-mails, letters, telephone calls, discussion sessions, and so forth.

Task Area 11: Follow-Up

Following delivery of the final report, the TFA metaevaluation team stood ready to help the client and other stakeholders interpret and apply the findings. Such follow-up activity can be crucially important to (1) help the client strengthen an evaluation system; (2) assist the client with disseminating the metaevaluation findings; (3) help interested stakeholders use the metaevaluation findings appropriately and productively; (4) ensure that various report recipients do not misinterpret, misrepresent, or misapply the findings; and (5) document and assess the metaevaluation's impacts and cost-effectiveness. Metaevaluation follow-up procedures will not always be desired and funded by a client. Clearly, metaevaluators cannot be expected to remain available to address a client's follow-up needs after completing a metaevaluation if no prior agreement and no associated funds exist to make this feasible. We recommend that in the initial contracting process—especially in large, summatively oriented metaevaluations—the metaevaluator and client carefully consider the desirability and possibility of engaging the metaevaluator to provide services following delivery of the final report. Often such follow-up services can contribute importantly to a metaevaluation's impact, but this contribution is unlikely to occur if the client has not planned and budgeted for the metaevaluator's follow-up involvement.

Metaevaluation Arrangements and Procedures

Using the eleven metaevaluation tasks just discussed, we next look at some of the specific arrangements and procedures that have proved useful in eleven metaevaluations: two of personnel evaluation systems, six of program evaluations, one of a needs assessment system,

one of alternative theoretical approaches to evaluation, and one of a large-scale student assessment system.

One of the personnel evaluation–related metaevaluation examples focused on the system the U.S. Marine Corps (USMC) used prior to the mid-1990s to evaluate the performance of officers and enlisted personnel, and the other addressed the system that the Hawaii Department of Education had been using to evaluate Hawaii's public school teachers. The examples of metaevaluations of program evaluations were focused on an independent evaluation of the New York City school district's tryout of the Waterford Integrated Learning System's computer-assisted basic skills program for elementary school students (Finn, Stevens, Stufflebeam, & Walberg, 1997); an evaluation of programs at the Appalachia Regional Educational Laboratory; an evaluation of the Reader Focused Writing program for the Veterans Benefits Administration (Datta, 1999; Grasso, 1999; Stake & Davis, 1999); an evaluation of a national distance baccalaureate program in an island nation of Southeast Asia; a small-scale, modest formative metaevaluation of Michael Scriven's first goal-free evaluation (of an early childhood program in a southern California school district); and an independent, formative metaevaluation of a task force's plan to evaluate the Himachal Pradesh Aadhar Programme (in Hindi, *Aadhar* means "support") in India, a nationally funded program for developing the state's capacity to strengthen its primary schools. (The metaevaluation of the plan for evaluating India's Himachal Pradesh Aadhar Programme was funded during the past decade by Cambridge Education in England; conducted by Stufflebeam from his base in Michigan; and rendered as an independent, formative metaevaluation to help an India national government–sponsored, fourteen-member evaluation team strengthen its evaluation plan.) The metaevaluation needs assessment example was a metaevaluation conducted by the U.S. Army command in Europe to assess needs assessments it was using in the mid-1980s to plan and offer courses to soldiers based in Europe. The metaevaluation of theoretical approaches was Stufflebeam's assessment (2001b) of alternative program evaluation models. The metaevaluation of a large-scale assessment system focused on an attempt by the National Assessment Governing Board (NAGB) to set achievement levels on the National Assessment of Educational Progress (NAEP; Stufflebeam, Jaeger, & Scriven, 1992; Vinovskis, 1999).

Space limitations prohibit in-depth discussion of any of these cases. Instead, we cite them to highlight particular arrangements and procedures that proved useful in conducting the eleven tasks identified earlier and to illustrate the different types of metaevaluation. Our intent is to help readers consider arrangements and procedures that might aid in the conduct of the eleven metaevaluation tasks and, where feasible, to show alternative ways of approaching different tasks. Clearly, a metaevaluation's context is important in determining when a procedure is or is not applicable and likely to be effective. In discussing these procedures, we offer caveats and commentary to help readers maintain circumspection in considering whether, when, and how to adapt and apply the cited arrangements and procedures. Following a restatement of each of the eleven tasks, we discuss the arrangements and procedures that we judge to have been useful in the referenced metaevaluations.

Task 1: Staff the Metaevaluation with One or More Qualified Metaevaluators

Clients should engage metaevaluation teams whose members have the needed technical qualifications, content knowledge, and credibility. The members should be respected and trusted by the stakeholders. In setting up the metaevaluation team for the USMC, it was important to include persons with military personnel evaluation experience, as well as expertise in the different aspects of metaevaluation. The metaevaluation for the New York City school district's computer-assisted basic skills program included the perspectives of metaevaluators with experience in educational research, program evaluation, educational policy, and school district operations, as well as the perspectives of men, women, and minorities. This team also could have used additional perspectives representing school- and classroom-level operations, computer technology, and possibly other elements. Generally the metaevaluation team's leader should clarify the required work and involve the client and stakeholders in appointing a qualified team.

In some situations, the client can afford to employ only a single metaevaluator and should then engage the most credible, capable metaevaluator that can be found. For example, over a period of years, the Evaluation Center at Western Michigan University employed William Wiersma to conduct its annual metaevaluations of Appalachia Regional Educational Laboratory project evaluations. He met this need exceptionally well because of his stature as an eminent educator, researcher, and author of widely used educational research methodology textbooks. He was thoroughly familiar with professional standards for evaluation and educational measurement. His more than forty years of research on schools and teacher education programs equipped him to understand education at all levels and to relate effectively to teachers, school administrators, policymakers, researchers, teacher educators, parents, and students. Wiersma's impressive credentials give an indication of the characteristics one should seek for a "lone ranger" metaevaluator assignment. The published metaevaluations by Datta (1999) and Grasso (1999) of the Stake and Davis (1999) evaluation of the Reader Focused Writing program for the Veterans Benefits Administration also illustrate the engagement of single, credible metaevaluators. The government of India chose Stufflebeam to conduct metaevaluations of plans for four statewide evaluations of the nation's primary school reform program (including the metaevaluation of the Himachal Pradesh Aadhar Programme evaluation), particularly because the government had decided to apply the 1994 Joint Committee program evaluation standards and because Stufflebeam had led development of the original 1981 edition of those standards.

Even in the face of restricted resources and a need for an independent formative metaevaluation, an evaluator can sometimes obtain metaevaluation services from a colleague at little or no cost. For example, Stufflebeam's assessment of Scriven's goal-free evaluation referenced earlier involved a fee of only one hundred dollars. On reviewing Scriven's initial evaluation plan, Stufflebeam judged that Scriven's plan to observe mainly classroom activities would miss program effects occurring on the school's playground and elsewhere outside the school. Following a revision of the evaluation plan, the program's main effects were identified—and not in classrooms but on the playground.

Sometimes an evaluator cannot or need not engage even a single independent metaevaluator, especially when the target evaluation is internal, small scale, and informal. Even then, an evaluator usefully can self-assess and report on evaluation plans, operations, and reports, having compared them against pertinent professional principles and standards.

Task 2: Identify and Arrange to Interact with the Metaevaluation's Stakeholders

The metaevaluation for the USMC was instructive in regard to the identification and involvement of stakeholders. Prior to contracting for the metaevaluation, the USMC leadership established two stakeholder review panels and arranged for systematic interaction between them and the metaevaluators. On the executive-level panel were eleven generals, four colonels, the sergeant major of the USMC, and some other officers. The second-tier stakeholder review panel included about twenty representatives from different ranks of officers and enlisted personnel. In setting up and arranging for systematic inputs from these broadly representative panels, the USMC sought to ensure that the metaevaluation would be relevant and credible to, and informed by, marines at all levels.

A USMC management office scheduled monthly meetings between the metaevaluation team and each panel, with each meeting scheduled for at least two hours. The metaevaluation team was contractually required to deliver printed reports at least ten working days in advance of the meeting, and the panelists were expected to read and prepare to discuss the reports. Collectively these reports spanned all major tasks in the metaevaluation: selection of standards for judging the USMC's personnel evaluation system; plans and instruments for obtaining information; diagnoses of strengths and weaknesses in the current personnel evaluation system; assessments of alternative personnel evaluation systems used in business, industry, and six other military organizations; generation and evaluation of three alternative personnel evaluation systems; and a plan for operationalizing and testing the selected new personnel evaluation system. For both groups, a designated general officer presided over each meeting.

Every meeting began with a briefing by the metaevaluators using an overhead projector, with copies of the transparencies distributed to all persons present. A period of questions, answers, and discussion followed. In concluding each meeting, the presiding general officer asked each panelist to address a bottom-line question. This general then summarized the meeting's main outcomes. Subsequently an assigned officer prepared and distributed a report of the discussion and conclusions reached at the meeting. These meetings were highly substantive and productive, with one lasting more than five hours without a break. We judge these USMC panelists to have been exemplary of the demeanor and contributions of evaluation clients; they were consummately professional, substantively engaged, and ultimately well-informed decision makers. It is noteworthy that at stakeholder review panel meetings virtually all members of both panels had thoroughly read and marked up the advance reports, and they engaged productively in in-depth discussion of the reported metaevaluation processes and findings.

A limitation was that the stakeholder review panels were top heavy with high-ranking officers, a significant issue considering that they had all been promoted by the personnel

evaluation system under investigation. Also, all of the panelists worked in the Washington, DC, area, not, for example, in California, Hawaii, Montana, Russia, Saipan, or Okinawa. There was a risk that voices and concerns of rank-and-file members throughout the USMC would not be sufficiently represented and heard. The metaevaluators had to strive mightily to convince the Washington-area generals of the need for additional inputs from outside the Washington, DC, area and from the full range of marine ranks. With these other inputs secured through surveys and site visits, the stakeholder involvement aspect of this metaevaluation improved.

We are confident that the structure involved in this project for the USMC could be beneficially applied in metaevaluations set in school districts, foundations, businesses, and other nonmilitary settings. Nevertheless, this example's stakeholder involvement procedures were largely dictated by the culture of the USMC and the fact that the commandant had mandated the metaevaluation and associated reform of the USMC's performance evaluation system. In other less structured institutions, a more flexible process of identifying and interacting with stakeholders might be preferable. Also, metaevaluators should keep in mind that some important metaevaluation stakeholders are identifiable only as the metaevaluation unfolds. In such cases, a metaevaluator and client should consider keeping open the question of who should be involved and informed throughout the metaevaluation and engaging these individuals over time as appropriate. To pursue the most effective process of interaction, a metaevaluator should carefully study and take into account the metaevaluation's context and the client organization's culture and preferred style of communication and involvement.

Not all metaevaluations need heavy involvement of stakeholders. As an example, there was minimal involvement of stakeholders in Stufflebeam's metaevaluation (2001b) of alternative theoretical approaches to evaluation. He engaged the authors of a number of the approaches evaluated to react critically to his characterizations and assessments of their approaches, obtained critiques of the draft manuscript from colleagues, and had an extensive exchange with Gary Henry, who was coediting the *New Directions for Evaluation* volume in which the metaevaluation report, titled *Evaluation Models,* eventually appeared (Stufflebeam, 2001b). Although some metaevaluations will require extensive, somewhat formal interaction with stakeholders, others require little, if any, interaction with either a narrow or a wide range of stakeholders. Also, sometimes systematic involvement of stakeholders throughout the metaevaluation process is not feasible. For example, Stufflebeam managed only limited interaction with stakeholders in his Michigan-based metaevaluations of India's primary school reform program evaluation plans and of the distance baccalaureate program in Southeast Asia. Nevertheless, in each of those cases one intensive, face-to-face exchange was arranged near the end of the metaevaluation process. In the distance baccalaureate case, Stufflebeam traveled to Southeast Asia, where he met with stakeholders and also gathered information beyond what had been mailed to him. In the India case, members of the evaluation task force traveled to Michigan and met for an extensive exchange of information with Stufflebeam. A metaevaluator should carefully consider a study's setting and exercise judgment in deciding how best to involve stakeholders.

Task 3: Agree on Standards, Principles, or Criteria to Judge the Evaluation System or Particular Evaluation

Evaluation is a professional activity. Therefore it is often appropriate and helpful to judge evaluations against the professional standards and principles of the evaluation field. Indeed, the need to invoke professional standards for evaluations is one of this book's key themes. The likelihood of harmonious conduct of an evaluation and securing its intended uses are enhanced when metaevaluators and their clients reach a clear advance understanding of the standards, principles, or criteria to be applied in evaluating a target evaluation. Depending on the particular situation, an evaluator and client may choose from among a range of published standards and principles pertaining to evaluation. Some examples follow, and, of course, Chapter 3 deals in depth with standards and principles for evaluations.

The American Educational Research Association, American Psychological Association, and National Council on Measurement in Education's *Standards for Educational and Psychological Testing* (1999) is especially useful for assessing educational testing programs (for example, NAGB's attempt to set achievement levels on NAEP and various state educational testing programs) and particular assessment devices. Applications of these standards to metaevaluate measurement devices are seen in the various volumes of the Buros Institute's Mental Measurements Yearbooks. Other potentially useful standards include the 2007 GAO government auditing standards and 2004 AEA guiding principles for evaluators.

The standards most used in our metaevaluations are the Joint Committee's personnel, program, and student evaluation standards (1981, 1988, 1994, 2003, 2009, 2011). They have been applied widely in North American educational evaluations. For example, the Hawaii State Board of Education adopted the Joint Committee program and personnel evaluation standards as state policy, stipulating that these standards be used to assess and strengthen Hawaii's system of educational accountability.

Although the Joint Committee standards were developed for use in evaluating North American educational evaluations, certain groups have found them appropriate and useful in other areas. For example, with minor modifications, the USMC adopted the 1988 Joint Committee personnel evaluation standards for use in assessing and reforming its system for evaluating officers and enlisted personnel. Similarly, General Motors (Orris, 1989) used the 1988 Joint Committee personnel evaluation standards to evaluate its system for evaluating executives. The U.S. Army applied the 1981 Joint Committee program evaluation standards to evaluate needs assessments conducted to help determine what courses the army should provide for soldiers stationed in Europe. The program evaluation standards were also used to evaluate the local, primary evaluation of the distance baccalaureate program in an island nation of Southeast Asia and plans for four statewide evaluations of India's current national school reform initiative.

In metaevaluations of the Stake and Davis (1999) evaluation, Datta (1999) employed the AEA guiding principles for evaluators (Shadish, Newman, Scheirer, & Wye, 1995a, 1995b), and Grasso (1999) mainly applied the AEA guiding principles (Shadish et al., 1995a, 1995b) but also referenced the 1994 Joint Committee program evaluation standards.

We acknowledge that a metaevaluator and a client evaluator (that is, an evaluator whose evaluation the metaevaluator is assessing) need not always reach an advance agreement on an explicit set of standards, principles, or criteria for judging an evaluation. Such formal negotiation of the bases for judging an evaluation tend to be unnecessary when only the evaluator needs the feedback, the orientation is formative rather than summative, the target is a draft evaluation plan or a particular issue in the evaluation, and the need for feedback is immediate. In such cases, an experienced metaevaluator's professional judgment may suffice. Metaevaluators should formally invoke pertinent evaluation standards, principles, and criteria when a metaevaluation would thereby be strengthened, when doing so is feasible, and especially when it is essential to maximize a metaevaluation's credibility. These conditions will usually pertain in summative metaevaluations and often in formative metaevaluations of fairly broad scope and large size.

Our experience with and the research on metaevaluation are too limited to yield definitive advice on weighting different standards for judging evaluations (see also Wingate, 2009). In general, we advise metaevaluators to begin by assuming that all the involved standards should be accorded equal importance. Following deliberation with stakeholders and careful thinking about a particular metaevaluation, one should, if appropriate, differentially weight standards or, especially, categories of standards. Sometimes it will be clear that some standards are not applicable in a particular metaevaluation. For example, the U.S. Army command in Europe decided that the Joint Committee's standards for accuracy, feasibility, and utility were highly applicable for judging the targeted needs assessment system, but that there was no need to invoke the propriety standards. If more were known about this case, one might justifiably disagree with the U.S. Army command's decision to exclude the propriety standards. Realistically, metaevaluators and their clients have to make choices about assigning relative importance to different standards or, as in the army case, categories of standards. In doing so, metaevaluators should carefully render judgments and document the basis for those judgments. In general, fewer standards will be applicable and important to the extent that a metaevaluation is small, formative, and directed only to an evaluator or a small audience.

Large-scale, summative metaevaluations employing the Joint Committee's standards often require that all the standards be applied. In such situations, a metaevaluator might justifiably decide that certain standards are so important that a failing grade on any of them would cause the evaluation to fail, even though high marks might have been attained on the other standards. In the case involving evaluation of alternative evaluation models, the metaevaluator (Stufflebeam) decided that no model should receive a passing grade if it failed any of the following standards: Service Orientation, Valid Information, Justified Conclusions, and Impartial Reporting. The stated rationale for this was that a model would be an unacceptable guide to evaluations if it did not help the evaluator assess a program's service to beneficiaries, answer an evaluation's questions, present defensible conclusions, and issue unbiased findings.

Task 4: Define the Metaevaluation Questions

In selecting questions for a metaevaluation, the fundamental considerations are to assess the evaluation for (1) how well it meets the requirements of a sound evaluation (merit) and (2)

its sufficiency in meeting the audience's evaluative information needs (worth). Fundamentally, a metaevaluator should assess the extent to which an evaluation conforms to professionally determined requirements of a sound evaluation, such as the AEA guiding principles or the Joint Committee standards. It follows that a metaevaluator should address a client group's particular questions.

Illustrating the last point, the National Assessment Governing Board, in concert with its contracted metaevaluators, defined more than twenty questions concerning its attempt to set achievement levels of basic, proficient, and advanced on the National Assessment of Educational Progress.

Two examples illustrating NAGB's specific questions are as follows:

- Is the membership of NAGB duly constituted, sufficiently representative of NAEP's constituencies, and effectively in touch with stakeholders so that it enjoys sufficient authority and credibility to set and secure use of achievement levels on NAEP?
- Are NAGB's policy framework and specifications for setting achievement levels sufficiently clear and consistent with the state of the relevant measurement technology to ensure that an appropriately representative group of formulators of standards can consistently and effectively set sound achievement levels on NAEP?

In general, a metaevaluator should ensure that a metaevaluation will determine the quality and overall value of a target evaluation or evaluation system and also address the audience's most important questions. Because some important metaevaluation questions may not be clear at the outset, a metaevaluator and client should consider the desirability of keeping open the possibility of identifying and addressing additional questions as the metaevaluation unfolds. Balance between following an evaluation's initial structure and remaining open to considering emergent questions is desirable.

Task 5: Issue a Memo of Understanding or Negotiate a Formal Metaevaluation Contract

As with most other evaluations, typically a metaevaluation should be grounded in a sound memorandum of agreement or formal contract. According to the Joint Committee (1994), evaluators and their clients should negotiate and document evaluation agreements that contain "a mutual understanding of the specified expectations and responsibilities of both the client and the evaluator" (p. 87). Such an agreement clarifies understandings and helps prevent misunderstandings between a client and metaevaluator, and it provides a basis for resolving any future disputes about a metaevaluation. Without this agreement, the metaevaluation process is vulnerable to misunderstandings, disputes, efforts to compromise the findings, attacks, or the client's withdrawal of cooperation and funds. As the Joint Committee (1994) further stated, "Having entered into such an agreement, both parties have an obligation to carry it out in a forthright manner or to renegotiate it. Neither party is obligated to honor decisions made unilaterally by the other" (p. 87). Written agreements for metaevaluations

should be explicit but should also allow for appropriate, mutually agreeable adjustments during a metaevaluation. Checklists by Stake (n.d.) and Stufflebeam (1999a) designed to help evaluators or metaevaluators and clients identify key contractual issues and make and record their agreements for conducting an evaluation or metaevaluation are available from www.wmich.edu/evalctr/checklists (also see Chapter 21). These checklists are designed to help evaluators or metaevaluators and their clients launch, stand by, and, as appropriate, modify agreements required to guide and govern an evaluation or metaevaluation.

In the metaevaluation of an evaluation of the Waterford Integrated Learning System project in the New York City school district (H. L. Miller, 1997), the metaevaluation team (Finn et al., 1997) was contracted not by the primary evaluators or the program directors but by an independent foundation. This helped the metaevaluators maintain their independence and issue sometimes unwelcome judgments without concern about having their contract canceled. Similarly, Cambridge Education in England served as the independent funding agent for Stufflebeam's metaevaluations of plans to evaluate four statewide school improvement programs in India. Many metaevaluations have the potential for conflict; when feasible, obtaining a contract and funds from a third party strengthens the metaevaluation's contractual grounding and viability.

An example of the hazards of proceeding without clear, advance written agreements is seen in the metaevaluation for NAGB (Stufflebeam, 2000a; Stufflebeam et al., 1992). Following completion of the contracted formative metaevaluation, the metaevaluators agreed to the client organization's urgent request, motivated by congressional pressure, for an immediate follow-up, summative metaevaluation. In view of the urgency of Congress's demand for a summative metaevaluation, the metaevaluators agreed to proceed with this follow-up metaevaluation before NAGB could formally process a contract through the federal bureaucracy. When NAGB subsequently was offended by the draft report and refused to pay for the summative metaevaluation work, there was no formal written agreement with which to press the issue. The lead metaevaluator's university never received payment for the summative metaevaluation work that had been agreed to informally but not in a written contract. The ensuing controversy over the draft summative metaevaluation findings stimulated a congressional investigation and almost resulted in cancellation of NAGB's funding. Much of this unfortunate controversy probably would have been avoided if the metaevaluators had insisted on reaching clear understandings and recording them in a signed contract before proceeding with the summative metaevaluation.

Task 6: Collect and Determine the Adequacy of Pertinent, Available Information

After agreeing on the terms to govern a metaevaluation, a metaevaluator needs to examine the target evaluation against pertinent evidence. Initially this involves collecting and assessing existing information. In some metaevaluations, this is the only information needed to reach the metaevaluative conclusions. Legitimate reasons for collecting additional information are that the existing information is technically inadequate, insufficient to address the full range of

metaevaluation questions, or not credible enough to earn the report recipients' confidence in the findings and conclusions. When the existing information is fully acceptable for producing a sound metaevaluation report, further data collection can be wasteful.

Datta (1999) and Grasso (1999) referenced Stake and Davis's published summary (1999) of their evaluation of the Reader Focused Writing program for the Veterans Benefits Administration as well as their full-length report. A key lesson for evaluators seen in these metaevaluations by Datta and Grasso is that evaluators and their clients can facilitate the conduct of metaevaluations by placing evaluation reports and supporting materials on a Web site for independent study and assessment of a completed evaluation. When negotiating the metaevaluation contract, metaevaluators and their clients are advised to consider making their study findings and procedures accessible on an appropriate Web site.

Stufflebeam's metaevaluation of the evaluation of the distance baccalaureate program is instructive concerning the kinds of extant information from which to begin a metaevaluation and how to handle that information. It was in the interest of administrators of the distance baccalaureate program to control the metaevaluation's costs, because travel from the United States to the other side of the world entails sizable expense. It was therefore agreed that the distance baccalaureate program administrators would send pertinent information to Kalamazoo about both the program and the external evaluation being conducted by evaluators at a university in the island nation. It was agreed that the metaevaluator would reference this material in reaching at least tentative judgments about the adequacy of the local evaluation of this program. A wide array of documents was sent to Kalamazoo: letters, plans, budgets, contracts, brochures, data collection forms, journal and newspaper articles, minutes of meetings, field notes, reports, and responses to reports. The metaevaluation's foci were the nature of the distance baccalaureate program, the background of the evaluation, evaluation plans and procedures, the evaluation process, the data, the conclusions, publicity for the program, and guidelines for the metaevaluation.

The metaevaluator's client emphasized that all judgments of the evaluation should be grounded in references to pertinent evidence. This, the client stated, would quell any stakeholder notions that the remotely conducted metaevaluation was only a set of vaguely informed opinions. Accordingly, the metaevaluator catalogued every piece of information used in the metaevaluation, giving its year of origination and a unique number within that year. In reporting a judgment for each of the thirty 1994 Joint Committee program evaluation standards, the metaevaluator referenced each catalogued information item used in reaching the judgment. Thus, the client and other stakeholders could review essentially all of the evidence used to reach the metaevaluative conclusions. In other metaevaluations, this documentation procedure has been useful not only for bolstering the metaevaluation report's credibility but also for maintaining a quite definitive history of the metaevaluation that can facilitate revisiting and studying the metaevaluation in later years, as might occur in doctoral dissertations.

Task 7: Collect New Information as Needed

Although the extant information for evaluating the evaluation of the distance baccalaureate program was substantial, the metaevaluator and client agreed that the information that had

been mailed to Kalamazoo was insufficient to generate and support metaevaluative conclusions. Thus, the metaevaluator traveled to the program's location to fill in some important information gaps. In addition to talking with the program's leaders and participating faculty, he also met with several students in the program and with leaders and faculty in the more traditional higher education programs. The additional information gathered led the metaevaluator to conclude that the assumed need—within the nation—for the distance baccalaureate program was questionable. Most of the nation's colleges and universities had many vacancies and were seeking students. Moreover, almost all of the program's students lived not in the remote areas of the country but in close proximity to institutions with openings for new students. Also, the quality of the program's offerings was highly variable, and there was unevenness in controlling exams to prevent cheating. The metaevaluator's findings were at variance with the highly positive local evaluation of the program. The metaevaluator concluded that the additional information obtained by making the on-site investigation proved essential to preparing and submitting a valid metaevaluation report. Such on-site investigation often is crucially important in metaevaluations, if for no other reason than to validate the information that has already been collected.

Another example of supplementing existing information with new information to reach metaevaluative conclusions occurred in Western Michigan University's assessment of Hawaii's teacher evaluation system. Jerry Horn first used extant information to judge Hawaii's system against each of the twenty-one 1988 Joint Committee personnel evaluation standards. He then supplemented this information with surveys of stratified random samples of Hawaii's public school teachers and administrators. The survey items were keyed to the twenty-one standards. The additional information not only corroborated the initial judgments but also provided an even stronger and more credible case that the existing teacher evaluation system was in serious need of reform.

Task 8: Analyze and Synthesize the Obtained Information

The wide array of information often used in metaevaluations requires a variety of quantitative and qualitative analysis procedures and tailored approaches to synthesis. In our metaevaluations, we have used line and bar graphs, pie charts, reanalysis of data from the target evaluation, and computer-assisted content analysis, among other analytic techniques. We have often let analysis results stand without converging them into a supercompressed grade or overall summary judgment. In these studies, we have usually presented separate results for an evaluation's utility, feasibility, propriety, and accuracy, followed by a narrative summary and discussion. As illustrated later in this chapter, however, sometimes we have combined scores on utility, feasibility, propriety, and accuracy into an overall score for an evaluation and also judged the overall evaluation to excellent, very good, good, fair, or poor.

In her metaevaluation of Stake and Davis's evaluation (1999) of the Reader Focused Writing program for the Veterans Benefits Administration, Datta (1999) employed a cross-break table to contrast the topics addressed in each of five case study reports. Based on this analysis, she observed, "Because there seemed to be only a few common elements reported on in each

site . . . the reliability in areas such as productivity seems uncertain. Sorting out idiosyncratic findings from incomplete inquiry is a bit difficult" (p. 350). In such a situation it would be a mistake to attempt a synthesis, which could only mask the idiosyncrasies of the different case study reports.

In a reanalysis of cost-effectiveness data for two alternative reading improvement programs (here, Program A and Program B), Stufflebeam arrived at conclusions that strongly contradicted the conclusions in the primary evaluation's draft report. (This analysis was part of the metaevaluation of the independent evaluation of the New York City school district's tryout of the Waterford Integrated Learning System's computer-assisted basic skills program for elementary school students [Finn et al., 1997].) That report had concluded that Program A was more cost effective than Program B, basically because Program A spent less than Program B on each student being served and because it was assumed that the two programs were equally effective for the served students. The assumption that the two programs were equally effective was not supported and was a basic flaw in the original analysis.

A key issue concerned what number of students should be included in the denominator used to determine each program's annual per pupil cost. Program A purported to serve every student in each participating school, and the evaluator had thus divided the sizable total annual program cost for each school by the number of students in the school. This analysis yielded a quite low per pupil cost for each school receiving Program A. Program B, which concentrated its reading recovery resources on students with substantial reading improvement needs, theoretically was less expensive and potentially more cost effective for a school as a whole than Program A. Also, Program B sought to serve each targeted student only until her or his reading proficiency was satisfactory and sustainable. For this program, the evaluator divided the total annual program cost for each school by the relatively small number of students this program served. Thus, each school's cost was high for each student served by Program B—about $8,000. On the basis of the different analyses used for the two programs, Program A's per pupil cost for each school was much lower than the per pupil cost for the schools served by Program B. The evaluator had gone on to suggest that Program A potentially was more cost effective than Program B, considering that the two programs were assumed to have comparable student achievement outcomes.

These conclusions were erroneous on both effectiveness and cost analysis grounds. Proponents of Program B, whose reading recovery resources were concentrated on only those students with reading deficiencies, argued that the program theoretically was less expensive and potentially more cost effective for a school as a whole than Program A. This was so because year after year, Program A spent a large amount of money but spread its services thinly across all students in the school. Program A thus extended remedial services to many students who did not need them and watered down the resources it might have concentrated on students with diagnosed reading deficiencies.

School district decision makers needed to know which program was more cost effective for a school as a whole in helping its students with reading deficiencies become good readers. A fairer cost analysis procedure would have divided each program's total annual cost by the total number of students in the involved school. This would have produced comparable school-wide

per pupil costs for each program. Over time, the evaluator might have assessed each program's cost-effectiveness by annually identifying the number of students in each school not requiring remediation in reading, then dividing this number into the school's annual expenditure for reading remediation. The lower the quotients over time, the greater would have been the indication that the program was attaining cost-effectiveness. Such a procedure could work provided that the student populations for schools receiving each program were comparable at the evaluation's outset. Although we do not have data to determine whether Program A or Program B would win in such a comparative study and analysis, we would bet on Program B. This is so because that hypothetical program would have concentrated its resources on students with assessed needs for remedial reading instruction and would have been designed to serve each targeted student intensively but only until his or her reading achievement was satisfactory and sustainable.

Task 9: Judge the Evaluation or Evaluation System in Terms of Its Adherence to Appropriate Standards, Principles, or Criteria

Following analysis and display of the obtained information, the evaluator should judge the target evaluation. Particularly important is the approach to judging the evaluation against the employed standards. Datta (1999) and Grasso (1999) basically keyed their narrative assessments of Stake and Davis's evaluation (1999) to an outline of the main standards in AEA's *Guiding Principles for Evaluators* (Shadish et al., 1995a, 1995b) and subparts of each.

Typically we have keyed our metaevaluations to each of the twenty-one personnel evaluation standards (Joint Committee, 1988) or the thirty program evaluation standards (Joint Committee, 1994, 2011) and, more specifically, to six to ten specific points (depending on what published checklist is used) associated with each standard. To support narrative judgments, we usually score the target evaluation on all points for each standard and then assign a predetermined scaled value meaning (for example, excellent, very good, good, fair, or poor) to the evaluation's adherence to each standard. Sometimes we subsequently have followed a set procedure to aggregate the scores across standards and produce judgments of the evaluation on each of the main requirements of utility, feasibility, propriety, and accuracy, and overall.

Tables 25.3 and 25.4 illustrate how a metaevaluation team based at Western Michigan University developed and presented judgments of the USMC's personnel evaluation system. The rubrics in Table 25.3, numbered 1 through 16, were used as rules for determining the degree to which the personnel evaluation system had satisfied standards in the four categories of utility, feasibility, propriety, and accuracy. All available relevant evidence was then used to identify and list the personnel evaluation system's strengths and weaknesses related to each standard in each category. Using these lists, judgments were formed about whether the system met, partially met, or failed to meet the standards in each category.

To summarize the results, the rubrics from Table 25.3 were used to prepare the summary matrix in Table 25.4. Basing its decision heavily on this analysis, the USMC decided to replace its personnel evaluation system with one that would better meet the standards.

Table 25.3 Sixteen Rubrics Used to Determine Whether a Military Branch's Personnel Evaluation System Satisfied an Evaluation's Requirements for Utility, Feasibility, Propriety, and Accuracy

	Degree of Fulfillment of Requirements		
Category of Standards	**Not Met**	**Partially Met**	**Met**
Utility	1. Three or more standards are not met.	2. At least three of the five standards are met or partially met, and at least one standard is not met. Or 3. Fewer than four standards are met, all five are either met or partially met, and no standard is unmet.	4. At least four or five standards are met, and none is unmet.
Feasibility	5. Three or four standards are not met.	6. At least two of the four standards are met or partially met, and at least one standard is not met. Or 7. Fewer than two standards are met, and no standard is unmet.	8. At least two of the four standards are met, and none is unmet.
Propriety	9. Three or more standards are not met.	10. At least three of the five standards are met or partially met, and one or two standards are not met. Or 11. Fewer than four standards are either met or partially met, and no standard is unmet.	12. At least four of the five standards are met, and none is unmet.
Accuracy	13. Four or more standards are not met.	14. At least five of the eight standards are met or partially met, and at least one standard is not met. Or 15. Fewer than five standards are met, at least four are either met or partially met, and no standard is unmet.	16. At least five of the eight standards are met, and none is unmet.

Note: This form is designed for use in judging the overall utility, propriety, feasibility, and accuracy of an evaluation. Use of this form's sixteen rubrics as decision rules requires that the user first judge whether the evaluation meets, partially meets, or fails to meet the detailed requirements of each of the twenty-one standards as they appear in Joint Committee (1988) and an additional utility standard (Transition to the New PRS [performance review system]) developed for this project. In some cases, a standard appropriately may be judged as not applicable, and such standards would have no impact on determining which of the rubrics fits the pattern of judgments.

Table 25.4 Conclusions on the Degree to Which a Military Branch's Personnel Evaluation System Satisfied Standards of Utility, Feasibility, Propriety, and Accuracy

Category of Standards	Conclusion	Rubric Used to Reach the Conclusion
Utility	Not met	1. Three or more standards are not met.
Feasibility	Partially met	6. At least two of the four standards are met or partially met, and at least one standard is not met.
Propriety	Partially met	10. At least three of the five standards are met or partially met, and one or two standards are not met.
Accuracy	Not met	13. Four or more standards are not met.

Task 10: Convey the Findings Through Reports, Correspondence, Oral Presentations, Workshops, and Other Means

Throughout most metaevaluations, there are important occasions for preparing and submitting metaevaluation reports. Typical reports are an initial metaevaluation plan, interim reports keyed

to the evaluation's important aspects, the final report, an executive summary, and a technical appendix or separate technical report. For each of these reports, it is usually advisable to prepare and submit a draft, follow this up with a meeting designed to orally communicate and discuss the draft report, and subsequently complete and submit the finalized version of the report (see Chapter 24). The core contents of the reports can be keyed to the guiding standards and principles for evaluations. Tables 25.3 and 25.4 illustrate ways to display the rationale for and results of standards-based judgments of evaluations. In delivering metaevaluation results, it is often appropriate to provide an executive summary, a set of supporting slides, a full-length report, and a separate technical report. Depending on advance agreements, it may also be appropriate to post the final metaevaluation report on a Web site, submit an executive summary for publication in a professional journal, and/or deliver oral presentations of the metaevaluation's implementation and findings.

Task 11: As Appropriate and Feasible, Help the Client and Other Stakeholders Interpret and Apply the Findings

Throughout the metaevaluation process, it is desirable for the metaevaluator to have regular, periodic exchanges with representatives of the key audiences. The evaluation for the USMC is illustrative of an extensive, functional, tightly scheduled reporting process. As noted previously, at the beginning of the metaevaluation, the USMC established two stakeholder review panels and issued specifications and a strict schedule for delivering reports. The reporting deadlines were closely linked to the USMC's need to make decisions for reforming its personnel evaluation system and to do so promptly. A key determinant of the deadline for delivering the final metaevaluation report in six months was the commandant's mandate that a plan for reforming the personnel evaluation system be submitted and approved before his fast-approaching retirement from his position. This constraint had serious implications for the metaevaluation's work schedule and the level of needed resources. Under the study's severe time limit, it was especially crucial that the metaevaluators and USMC leaders met regularly and often to review findings and apply them to the ongoing process of reforming the personnel evaluation system.

The referenced experience with the Hawaii Department of Education was similar in some respects to the experience with the USMC. At the metaevaluation's outset, the department appointed a review panel that represented the various interests in the state's public education system. The panel included a broadly representative group of persons with interests in the state's teacher evaluation system. The panel was chaired by the state superintendent of public instruction, and its members included teachers, school administrators, parents, students, the head of Hawaii's teachers' union, the head of Hawaii's association of school administrators, the majority leaders of Hawaii's state senate and house of representatives, the chairman of the Hawaii Board of Education, an officer stationed at a U.S. military base in Hawaii, and representatives of local business and industry. The metaevaluation team and members of the Hawaii Department of Education met regularly with this group to discuss and obtain inputs concerning the ongoing metaevaluation of Hawaii's teacher evaluation system. The panel was

asked to review, critique, and discuss metaevaluation plans, tools, and draft reports; facilitate implementation of the metaevaluation; and help the department use the findings. The review panel helped clarify the questions, provided valuable critiques of survey instruments and draft reports, and used the findings to generate recommendations for improving the state system of teacher evaluation. By being involved in the metaevaluation process, the review panel developed ownership of the findings and became a powerful, informed resource for helping to chart and obtain support for the needed reforms.

Review Panels Versus Advisory Panels

It was important that this group was labeled a "review panel" and not an "advisory panel." The orientation was that the panel members were qualified to critique draft plans, schedules, interview protocols, and reports from their own perspective; comment on such matters as feasibility and clarity; and facilitate data collection. The panel members were not necessarily qualified to provide technical advice for improving such metaevaluation aspects as the design, instruments, and analysis procedures. In our experience, reference groups sometimes become dysfunctional and counterproductive when they are accorded an aura of unmerited expertise by virtue of being labeled an "advisory panel."

Parallel to the involvement of the review panel, the metaevaluators engaged educators in Hawaii to help carry out the metaevaluation. They were especially helpful in obtaining relevant documents, files, and data tapes and in arranging for other educators to participate in interviews, surveys, and focus group meetings.

Metaevaluation can and often should be a collaborative effort, especially when the aim is to help an organization assess and reform its evaluation system. When the aim is to protect the public from being misinformed by evaluations of specific entities, metaevaluators must maintain proper distance to ensure an independent perspective. Even then, however, metaevaluators should communicate appropriately with audiences for the metaevaluation reports to secure their confidence, interest, assistance, understanding, and informed use of findings.

Comparative Metaevaluations

Sometimes a metaevaluation involves a comparative assessment of a number of evaluations (also see T. D. Cook & Gruder, 1978). For example, professional societies such as AEA and the American Educational Research Association do so when they rate evaluations as a basis for making awards to outstanding evaluations. Figure 25.1 provides an example of how eight hypothetical candidate evaluations might be subjected to a comparative metaevaluation. The hypothetical evaluations are listed in order of judged merit. The ratings are in relation to the 1994 Joint Committee program evaluation standards and hypothetically have been derived by using a special checklist keyed to those standards.

Assume that each evaluation was rated on each of the thirty program evaluation standards by judging whether the study met each of ten key features of the standard (as defined in Stufflebeam's metaevaluation checklist [1999b]). Further assume that each evaluation was

then judged on each standard as follows: 9–10 = excellent, 7–8 = very good, 5–6 = good, 3–4 = fair, and 0–2 = poor. The score for each evaluation on each of the four categories of standards (utility, feasibility, propriety, accuracy) was then determined by summing the following products: 4 × the number of excellent ratings, 3 × the number of very good ratings, 2 × the number of good ratings, and 1 × the number of fair ratings.

Judgments of each evaluation's strength in satisfying each category of standards were subsequently determined according to percentages of possible quality points for the category of standards as follows: 93–100 percent = excellent, 68–92.99 percent = very good, 50–67.99 percent = good, 25–49.99 percent = fair, and 0–24.99 percent = poor. This was accomplished by converting each category score to the percentage of the maximum score for the category, then multiplying by 100. In Figure 25.1 the four equalized scores were next summed, divided by 4, and compared with the total maximum value of 100. The evaluation's overall merit was then judged as follows: 93–100 = excellent, 68–92.99 = very good, 50–67.99 = good, 25–49.99 = fair, and 0–24.99 = poor. This procedure unequally weights different standards in the process of computing a total score and overall rating. This is because the four categories contain unequal numbers of standards, and the individual standards in categories with fewer standards have more weight than the individual standards in categories with more standards. An alternative means of determining a total score and overall rating is to sum and average the thirty individual standard scores. We advise metaevaluators to compute, assess, and discuss the extent of agreement between the calculations derived using both the total score and the overall rating approaches.

Regardless of each evaluation's total score and overall rating, we would judge any evaluation as failed if it received a poor rating on the vital standards of P1 Service Orientation, A5 Valid Information, A10 Justified Conclusions, and A11 Impartial Reporting.

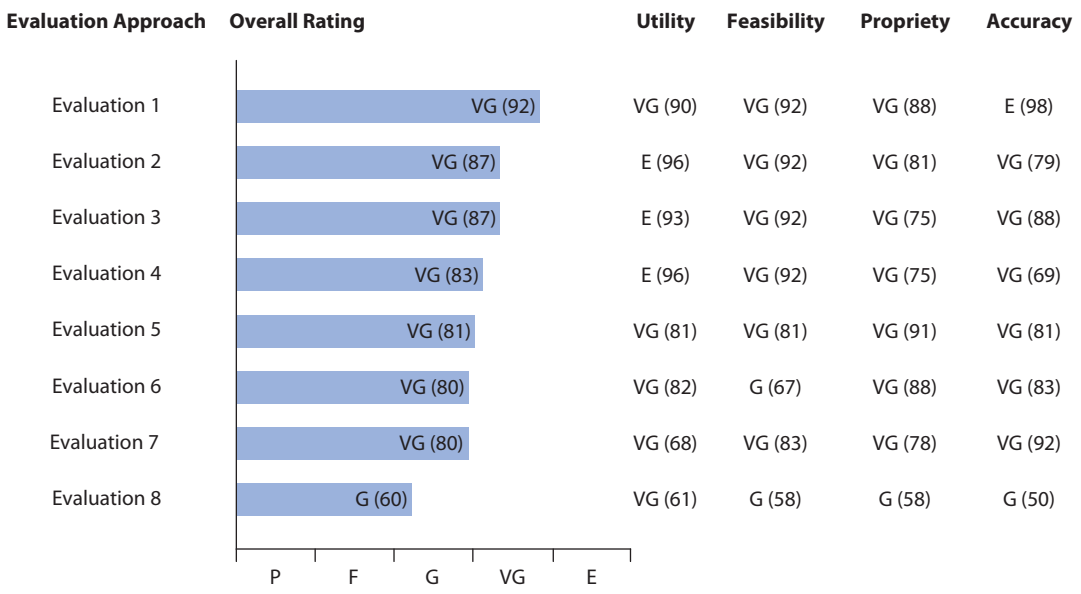

Figure 25.1 Ratings of Candidate Program Evaluations
Note: P = poor, F = fair, G = good, VG = very good, E = excellent.

Checklists for Use in Metaevaluations

Checklists can be useful in metaevaluations. The Evaluation Contracts Checklist (Stufflebeam, 1999a) and the Program Evaluations Metaevaluation Checklist (Stufflebeam, 1999b, 2011b) plus additional checklists may be accessed at www.wmich.edu/evalctr/checklists. Included in that repository are checklists designed for use in evaluating personnel, programs, and materials. They are applicable to the conduct of primary evaluations and metaevaluations. Among others, these checklists include Scriven's Key Evaluation Checklist (2007), a checklist by House and Howe (2000b) for guiding and assessing deliberative democratic evaluations, Shepard's Checklist for Evaluating Large-Scale Assessment Programs (1977), and one on the AEA guiding principles for evaluators (Stufflebeam, Goodyear, Marquart, & Johnson, 2005). (For a general discussion of evaluation checklists, see Stufflebeam, 2001a.)

The Role of Context and Resource Constraints

The preceding discussion of metaevaluation must be tempered by considerations of the reality constraints in evaluation work. It will not always be important or feasible to do a formal metaevaluation. In particular, the client may be unwilling to commission and fund an independent metaevaluation. All the cases referenced in this chapter are examples where a client requested and funded a metaevaluation. Even then, the cases varied considerably in regard to the client's need for extensive, formal feedback and the size of the budget. The amounts of money invested generally were in the range of $10,000 to $30,000, but the smallest metaevaluation cost only $100, and the metaevaluation for the USMC was funded at about $446,000. Generally the small, formative evaluations required much less money and effort for metaevaluation than did the large-scale, summative evaluations. Usually cost should not be a deterrent to obtaining some level of metaevaluation. Typically the size of budgets for metaevaluations is minuscule compared with the cost of the target evaluation—often less than 2 percent of the target evaluation's budget. Moreover, in large-scale, high-stakes evaluations, metaevaluations often can be judged cost-free when their costs are compared with the value of the benefits they produce (Scriven, 1991).

Nevertheless, sometimes evaluators or clients of an evaluation will not request or need a formal metaevaluation. For example, a formal metaevaluation might not be needed for an evaluation system that was subjected to a metaevaluation relatively recently and that subsequently has operated relatively free of complaints and observed problems. Also, individual personnel evaluations typically require no metaevaluations, except when one is triggered by an appeal of the findings. Many government agencies, accrediting organizations, and charitable foundations seek no metaevaluations of the evaluations they sponsor, presumably because they trust their system of monitoring and oversight (although such trust is not always justified). Very small-scale, formative evaluations, as when one evaluates a small project or a course for purposes of improvement, might not need or be amenable to any kind of formal metaevaluation.

Although there are evaluations that require little or no metaevaluation, it is always appropriate for an evaluator to plan and carry out even small-scale formative evaluations with a metaevaluation mind-set. One of the best ways to do this is to thoroughly study and internalize the key messages of the Joint Committee's standards (1981, 1988, 1994, 2003, 2011); the 2004 AEA guiding principles for evaluators; or other relevant standards. Having the underlying metaevaluation principles in mind is invaluable in planning an evaluation, dealing with issues and problems as they arise, advising evaluation participants in regard to the dilemmas they face, and—after the fact—taking stock of what the evaluation accomplished.

Summary

Metaevaluations serve all segments of society. They help ensure the integrity and credibility of evaluations and are thus important to both users and producers of evaluations. Metaevaluations often are needed to scrutinize evaluations of charitable services; research and development projects; equipment and technology; state assessment systems; new, expensive curricula; policies and strategic plans; automobiles and refrigerators; hospitals and other organizations; and engineering plans and projects. They also are needed to assess and help improve systems used to evaluate physicians, military officers, researchers, evaluators, public administrators, teachers, school principals, students, and others. In the case of appeals, metaevaluations are needed to assess the soundness and fairness of evaluations of individual employees. As seen in these examples, metaevaluations are in public, professional, institutional, and personal interests.

As professionals, evaluators themselves need to regularly subject their evaluation services to internal and independent review. Sound metaevaluations provide evaluators with a quality assurance mechanism they can use to examine and strengthen evaluation plans, evaluation operations, draft reports, and means of communicating findings. The prospect and fact of metaevaluations should help keep evaluators on their toes, push them to produce defensible evaluation services, and guide them to improve their services over time.

Metaevaluation is as important to the evaluation field as auditing is to the accounting field. Society would be seriously at risk if it depended only on accountants for its financial information, without acquiring the scrutiny of independent auditors. And parents, students, educators, government leaders, businesspersons, and consumers in general would be at risk to the extent that they cannot trust evaluation findings.

Despite the strong case that can be made for metaevaluation, not all evaluations require or merit a metaevaluation. Small-scale, locally focused, and improvement-oriented evaluations may not require any special metaevaluation. Making such determinations is a matter for careful judgment by the evaluator and client; they should take into account the local setting and especially the audience for the target evaluation. In deciding whether to commission or conduct a metaevaluation, the evaluator and client should keep in mind that a metaevaluation's cost is typically small compared to the cost of the target evaluation, and that the value of the metaevaluation's benefits can far outweigh the metaevaluation's costs.

We defined metaevaluation generally as the process of evaluating evaluations. We defined it operationally as the process of delineating, obtaining, and applying descriptive and judgmental information about an evaluation's utility, feasibility, propriety, accuracy, and accountability for the purposes of guiding the evaluation and reporting its strengths and weaknesses. Based on these definitions, we presented a general, eleven-task methodology for metaevaluation. We referenced example metaevaluations to identify key metaevaluation arrangements and techniques of use in carrying out each metaevaluation task. We defined needed qualifications for carrying out metaevaluations. We also discussed practical procedures and tools for conducting metaevaluations, including checklists for contracting for metaevaluations and judging evaluations, the process of contracting for metaevaluations with third parties, review panels, feedback sessions, and rubrics and analysis protocols for judging evaluations.

Undergirding this chapter is the strong recommendation that evaluators ground their metaevaluations in professional standards and principles for evaluations. For evaluators working in North America, we recommended the use of the 2004 AEA guiding principles for evaluators; the 2007 GAO government auditing standards; and the professional standards for evaluations of programs, personnel, and students issued by the Joint Committee (1981, 1988, 1994, 2003, 2011).

Evaluators are making progress in developing and applying methods for use in metaevaluations. Sustaining and increasing efforts to systematize and increase the rigor, relevance, and contributions of metaevaluations are in the interest of professionalizing the evaluation field and serving society well.

REVIEW QUESTIONS

1. List at least five reasons an evaluator could give a client to justify the expense of contracting for an independent metaevaluation of an evaluation system, such as a state's teacher evaluation system.
2. Compare and contrast the concepts of metaevaluation and meta-analysis.
3. In general, what steps would you follow to appropriately apply a standards-based checklist for reaching judgments of an evaluation's utility, feasibility, propriety, accuracy, and overall soundness?
4. What is this chapter's operational definition of metaevaluation? Summarize how you would explain this definition to a colleague who is unfamiliar with metaevaluation.
5. According to this chapter, what are the eleven main tasks of a metaevaluation?
6. What advantages does an evaluator or client attain by contracting with a third party to conduct a particular metaevaluation?
7. What are possible reasons why evaluators in a totalitarian country might choose not to adopt and apply any of the Joint Committee's sets of standards?

8. Under what circumstances is it likely to be inappropriate or unnecessary to conduct a metaevaluation of a particular evaluation?
9. What is the role of stakeholder review panels in metaevaluations, and what has been our argument against referring to such groups as "advisory panels"?
10. List at least ten things that can go wrong in an evaluation, and discuss how a sound metaevaluation could be designed to prevent or expose these.

Group Exercises

Exercise 1

Compare and contrast formative metaevaluation and summative metaevaluation in terms of purpose, timing, who should conduct the metaevaluation, and audiences.

Exercise 2

Discuss the proposition that formative and summative metaevaluations should be conducted internally as well as independently.

Exercise 3

Ask one member of your group to identify and bring to the group's next meeting a description of an evaluation for group members to use in applying what they have learned about metaevaluation. At your next group meeting,

1. List at least eight features of this evaluation that should be assessed in a metaevaluation.
2. Project main metaevaluation questions to be answered.
3. Select standards for assessing this evaluation.
4. Decide on the procedures your group would use to answer the metaevaluation questions and to judge the evaluation against the chosen standards.
5. Outline the executive summary of the projected metaevaluation report.

Exercise 4

Visit www.wmich.edu/evalctr/checklists, and review the checklists your group might use to conduct a metaevaluation of the evaluation discussed in exercise 3. Which one or more checklists would your group choose to conduct the metaevaluation? Justify your choice.

Exercise 5

This chapter has listed skills required by a competent metaevaluator. Discuss these skills with your group. As a group, list skills in which at least some members feel especially well qualified.

Subsequently, list skills that some or all members would like to acquire or strengthen. Finally, for any identified areas of deficiency, discuss actions that could be taken to attain the needed skills.

Note

1. This chapter is an amalgam and update of Stufflebeam's *Meta-Evaluation* (1974), "Meta-Evaluation" (2011a), "Meta-Evaluation: An Overview" (1978), "The Methodology of Metaevaluation" (2000b), and "The Metaevaluation Imperative" (2001c). It also draws on Chapter 27 of this book's first edition (Stufflebeam and Shinkfield, 2007).

Suggested Supplemental Readings

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Evaluation Association. (2004). *Guiding principles for evaluators*. Washington, DC: Author. Retrieved from http://www.archive.eval.org/Publications/GuidingPrinciples.asp

Beywl, W. (2000). Standards for evaluation: On the way to guiding principles in German evaluation. In C. Russon (Ed.), *The program evaluation standards in international settings* (pp. 60–67). Kalamazoo: Western Michigan University, Evaluation Center.

Cooksy, L. J., & Caracelli, V. J. (2005). Quality, context, and use: Issues in achieving the goals of metaevaluation. *American Journal of Evaluation*, *26*, 31–42.

Cooksy, L. J., & Caracelli, V. J. (2009). Metaevaluation in practice: Selection and application of criteria. *Journal of MultiDisciplinary Evaluation*, *6*(11), 1–15.

Datta, L.-E. (1999). CIRCE's demonstration of a close-to-ideal evaluation in a less-than-ideal world. *American Journal of Evaluation*, *20*, 345–354.

Finn, C. E., Stevens, F. I., Stufflebeam, D. L., & Walberg, H. J. (1997). A meta-evaluation. *International Journal of Educational Research*, *27*, 159–174.

Grasso, P. G. (1999). Meta-evaluation of an evaluation of Reader Focused Writing for the Veterans Benefits Administration. *American Journal of Evaluation*, *20*, 355–371.

Jang, S. (2000). The appropriateness of Joint Committee standards in non-Western settings: A case study of South Korea. In C. Russon (Ed.), *The program evaluation standards in international settings* (pp. 41–59). Kalamazoo: Western Michigan University, Evaluation Center.

Joint Committee on Standards for Educational Evaluation. (1981). *Standards for evaluations of educational programs, projects, and materials*. New York, NY: McGraw-Hill.

Joint Committee on Standards for Educational Evaluation. (1988). *The personnel evaluation standards*. Thousand Oaks, CA: Corwin Press.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Joint Committee on Standards for Educational Evaluation. (2003). *The student evaluation standards*. Thousand Oaks, CA: Corwin Press.

Joint Committee on Standards for Educational Evaluation. (2009). *The personnel evaluation standards: How to assess systems for evaluating educators* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Joint Committee on Standards for Educational Evaluation. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Miller, H. L. (Ed.). (1997). The New York City Public Schools integrated learning systems project: Evaluation and meta-evaluation [Special issue]. *International Journal of Educational Research*, *27*(2).

Orris, M. J. (1989). *Industrial applicability of the Joint Committee's personnel evaluation standards*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Sanders, J. R. (1995). Standards and principles. In W. R. Shadish, D. L. Newman, M. A. Scheirer, & C. Wye (Eds.), *Guiding principles for evaluators* (pp. 47–52). New Directions for Program Evaluation, no. 66. San Francisco, CA: Jossey-Bass.

Scriven, M. (1969). An introduction to meta-evaluation. *Educational Products Report*, *2*(5), 36–38.

Scriven, M. (1994). Product evaluation: The state of the art. *Evaluation Practice*, *15*, 45–62.

Scriven, M. (2000). *The logic and methodology of checklists*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/papers/logic&methodology_dec07.pdf

Smith, N. L., Chircop, S., & Mukherjee, P. (2000). Considerations on the development of culturally relevant evaluation standards. In C. Russon (Ed.), *The program evaluation standards in international settings* (pp. 29–40). Kalamazoo: Western Michigan University, Evaluation Center.

Stake, R. E., & Davis, R. (1999). Summary evaluation of Reader Focused Writing for the Veterans Benefits Administration. *American Journal of Evaluation*, *20*, 323–344.

Stufflebeam, D. L. (1974). *Meta-evaluation* (Occasional Paper Series, Paper #3). Kalamazoo: Western Michigan University, Evaluation Center.

Stufflebeam, D. L. (1978). Metaevaluation: An overview. *Evaluation & the Health Professions*, *1*(2), 146–163.

Stufflebeam, D. L. (2000). *Guidelines for developing evaluation checklists: The checklist development checklist (CDC)*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/guidelines_cdc.pdf

Stufflebeam, D. L. (2000). Lessons in contracting for evaluations. *American Journal of Evaluation*, *21*, 293–314.

Stufflebeam, D. L. (2000). The methodology of metaevaluation. In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed., pp. 457–496). Norwell, MA: Kluwer.

Stufflebeam, D. L. (2001). Evaluation checklists: Practical tools for guiding and judging evaluations. *American Journal of Evaluation*, *22*, 71–79.

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

Stufflebeam, D. L. (2001). The metaevaluation imperative. *American Journal of Evaluation*, *22*, 183–209.

Stufflebeam, D. L., Jaeger, R. M., & Scriven, M. (1992, April). *A retrospective analysis of a summative evaluation of NAGB's pilot project to set achievement levels on the National Assessment of Educational Progress*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Taut, S. (2000). Cross-cultural transferability of the program evaluation standards. In C. Russon (Ed.), *The program evaluation standards in international settings* (pp. 5–28). Kalamazoo: Western Michigan University, Evaluation Center.

U.S. Government Accountability Office. (2003). *Government auditing standards* (GAO-03-763G). Washington, DC: U.S. Government Printing Office.

U.S. Government Accountability Office. (2007). *Government auditing standards* (GAO-07-731G). Washington, DC: U.S. Government Printing Office.

Vinovskis, M. (1999). *Overseeing the nation's report card: The creation and evolution of the National Assessment Governing Board (NAGB)*. Washington, DC: National Assessment Governing Board.

Widmer, T., Landert, C., & Bacmann, N. (2000). Evaluation standards recommended by the Swiss Evaluation Society (SEVAL). In C. Russon (Ed.), *The program evaluation standards in international settings* (pp. 81–102). Kalamazoo: Western Michigan University, Evaluation Center.

CHAPTER 26

INSTITUTIONALIZING AND MAINSTREAMING EVALUATION

We begin this concluding chapter by summarizing the book's nine themes. We subsequently address the practical culmination of these themes by discussing and illustrating steps organizations can take to institutionalize and mainstream systematic practices of program evaluation. We present this chapter's conclusions and recommendations as our best judgments of how an organization can proceed to design and install a sound system of evaluation. This chapter reflects the extensively documented material in previous chapters but includes only those citations of practical use in institutionalizing and mainstreaming evaluation.

Review of this Book's Themes

This book's orientation is both theoretical and practical. These two factors are intertwined throughout the book and evident in nine themes. The first two themes are that (1) the evaluation discipline should be grounded in sound theory—that is, a coherent set of conceptual, hypothetical, pragmatic, and ethical principles forming a general framework to guide the study and practice of evaluation, and (2) society needs and is using evaluations to inform decisions and hold service providers accountable for the implementation and outcomes of the services they provide. To the ends of informed decision making and accountable programs, evaluators are developing and deploying a responsive, distinctive evaluation methodology. In Part One we presented evaluation's fundamentals by discussing the evaluation field's history and current state, the nature of evaluation theory, and professional standards for evaluations.

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **A rationale and key principles for, and definitions of, institutionalizing and mainstreaming evaluation**
- **Early efforts to help organizations institutionalize and mainstream evaluation**
- **Advances in evaluation that are of use in institutionalizing and mainstreaming evaluation**
- **A checklist for use in establishing a unified organizational evaluation system**

Part Two was a consumer report examination of twenty-three approaches often used to evaluate programs, including six illicit and seventeen defensible approaches. That part's assessments were keyed to the book's third, fourth, and fifth themes: (3) evaluators and clients must guard against the use of unsound, often corrupt inquiry approaches that masquerade as sound evaluation but, in fact, are designed to mislead right-to-know audiences or prevent some of them from obtaining evaluation findings; (4) evaluators can choose from a range of defensible evaluation approaches; and (5) evaluators should employ professional standards to assess and select evaluation approaches and ensure the quality of particular evaluations.

Part Three extended the development of these three themes by providing details, procedures, and examples of application for six especially noteworthy evaluation approaches: experimental and quasi-experimental design evaluation; case study evaluation; the context, input, process, and product (CIPP) model for evaluation; consumer-oriented evaluation; responsive or stakeholder-centered evaluation; and utilization-focused evaluation.

Part Four addressed the book's sixth and seventh themes: (6) evaluators should employ systematic procedures that possess general applicability across evaluation approaches and provide sound protocols for proceeding through an evaluation's start-up, design, budgeting, contracting, information collection, analysis, synthesis, reporting, and follow-up stages, and (7) evaluators should involve stakeholders in the evaluation process to hear and consider their inputs and enhance prospects for their appropriate and beneficial use of findings.

Part Five provides a capstone for the book's previous chapters by addressing the eighth and ninth themes: (8) as professionals, evaluators must subject their evaluations to metaevaluation, and (9) organizations of all types should institutionalize and mainstream sound evaluation practices as a vital part of planning programs, conducting the programs, and meeting requirements for accountability, because at its core, every discipline and service area needs sound evaluation to confirm and continually strengthen its claim that it is effectively serving clients and the public interest as well as fulfilling other defensible purposes.

Overview of the Remainder of the Chapter

In the previous chapters we treated evaluation as more or less an ad hoc activity—one that is mounted in response to particular needs related to evaluative guidance, feedback, and conclusions. We therefore provided concepts and procedures that evaluators and their clients could learn about and embrace and then apply after having decided to pursue particular evaluations, especially project evaluations. This chapter takes the discussion of systematic evaluation to another level. It is addressed to organizations that can and should install and regularly operate an evaluation system for assessing, for example, ongoing programs as well as projects. The chapter is targeted to help institutional policy boards, administrators, and staff institutionalize a sound system of evaluation within their organization and mainstream uses of the evaluation system at all levels and in all important parts of the organization. The chapter emphasizes that organizational evaluation systems should involve basically all of an organization's staff

members in an organization's conduct and use of rigorous evaluations, especially for the purposes of decision making, improvement, and accountability.

By *institutionalizing evaluation,* we mean that an organization defines, installs, regularly operates, and uses results from an evaluation system that is relatively permanent in the organization. By *mainstreaming evaluation,* we mean the evaluation system functions at all levels of the organization by assessing all facets that are vital to fulfilling the organization's mission, and that the organization's full range of personnel are engaged in the conduct and use of evaluations.

Rationale and Key Principles for Institutionalizing and Mainstreaming Evaluation

The recommendation to institutionalize and mainstream evaluation is both reasonable and practical. Human beings evaluate in some way or other whenever they make decisions and also when they judge how well the execution of past decisions worked out. The problem is that many such evaluations are hurried, haphazard, impulsive, biased, and/or not grounded in sound information. Clearly, on the one hand, poor evaluations can lead to poor decisions and bad outcomes. On the other hand, sound evaluations can lead to success, credibility, and pride in one's quality of service and accomplishments. In any organization it is important to ensure as much as is practicable that decisions, expenditures, actions, and accountability reports are grounded in sound evaluations.

Although many organizations provide for and support systematic approaches to evaluation at top administrative levels, many do not spread and support systematic evaluation throughout an organization's lower levels. In this regard it is essential that systematic evaluation be considered the domain and responsibility not just of top administrators and assigned evaluation specialists, but of every professional member of the organization. Indeed, evaluating and being accountable for the quality of one's services is the essence of being a professional. Ideally, an organization should officially adopt, support, instrument, and apply systematic evaluation in all of its key programs and activities, and basically all staff should develop skills in and employ sound processes of evaluation designed to guide their thinking and planning and to ensure and help document the effectiveness of their actions.

It seems logical that a unified and fully functional evaluation system could be one of an organization's most effective tools. It would provide the organization's personnel with a common process and language for determining needs for evaluation, defining evaluation questions, negotiating evaluation agreements, choosing appropriate evaluation methods, defining and reporting to evaluation audiences, and holding evaluations to standards of sound evaluation. It would provide information on the full range of issues in the organization. Such issues might include, among others, client needs, service plans and budgets, program or service implementation and costs, organizational accomplishments, barriers to program effectiveness, and constituents' and community members' perceptions and support of the organization. A sound

organizational evaluation system would provide feedback throughout the year, not just at the end. It would be applied at all levels: policy and administration, departments, special projects, work groups, and individual staff members. It would give the organization and each staff member a shared, general approach for evaluating programs, projects, services, and so on. It would stress that evaluations should be designed and used for improvement and accountability purposes, and to maintain a historical record. It would aid, not hinder, progress at all levels of the organization.

Early Efforts to Help Organizations Institutionalize Evaluation

In 1970 Egon Guba and Daniel Stufflebeam collaborated on a paper titled *Strategies for the Institutionalization of the CIPP Evaluation Model*. They proposed that organizations install a unified evaluation approach by which not only to efficiently address the many accountability requirements then being pressed on them but also to regularly supply intended audiences with information for planning and guiding projects and other organizational activities. The paper emphasized that external evaluators, although important, could not fully serve organizations, because, among a number of key reasons, there were far too few evaluation centers, companies, and consultants to address the evaluation needs of the vast number of organizations being required to evaluate their programs.

According to Guba and Stufflebeam (1970), organizations needed an ongoing process of internal evaluation to help staff members constantly learn from their experiences and improve practices. Such a process would also enable organizations regularly to report accomplishments to sponsors and other external audiences. Moreover, internal evaluations would serve an organization better if all members subscribed to and employed the same sound view of evaluation.

Guba and Stufflebeam (1970) offered the CIPP model as a comprehensive framework of pertinent questions and information that diverse groups could use as a common evaluation philosophy and language. They then examined the model's requirements to determine what steps an organization would need to follow in developing and employing a fully functional evaluation system. In general, they advised organizations to empower themselves by institutionalizing a systematic process of evaluation that would inform the organization's ongoing decision-making process; help ensure its accountability to boards, funders, and constituents; and complement and help cross-check external evaluations of the organization's contributions.

Guba and Stufflebeam drew their evaluation recommendations from a range of instructive evaluation experiences. Together and independently they had been contracted to help a number of different types of organizations successfully apply their evaluation ideas. These organizations included school districts (for example, in Columbus, Xenia, and Cincinnati, Ohio; Saginaw, Lansing, and Detroit, Michigan; Dallas, Fort Worth, and Houston, Texas; and Des Moines, Iowa); regional educational laboratories (for example, the Austin, Texas, Southwest Educational Development Laboratory and the Northwest Regional Educational Laboratory in Oregon); state education departments (including in Montana, Michigan, and

Ohio); research and development centers (such as the National Center for Research on Vocational and Technical Education and the Wisconsin Research and Development Center on Student Learning); and the U.S. Office of Education.

Recent Advances of Use in Institutionalizing and Mainstreaming Evaluation

In their 1970 paper, Guba and Stufflebeam could draw on only their own experiences and modest developments within the evaluation field, compared to which the evaluation experiences and progress discussed in this book are vast. The more recent relevant developments include

- Professional societies of evaluators and their attendant professional journals
- Professional standards for program evaluation, personnel evaluation, and student evaluation (discussed in detail in Chapter 3)
- A large and growing evaluation literature
- Evaluation degree programs and a small but steady supply of well-trained graduates (Coryn, Stufflebeam, Davidson, & Scriven, 2010; LaVelle & Donaldson, 2010)
- Organizations that have successfully institutionalized an evaluation system, such as the U.S. Congress through its Government Accountability Office; the Jefferson County, Kentucky, and Dallas, Texas, independent school districts, among many others; and the World Bank
- Growing experience in conducting metaevaluations (see Chapter 25)
- Alternative evaluation models that have been widely applied (for particular examples, see Chapters 11 through 16)
- Consensus that quantitative and qualitative approaches are complementary, not contradictory
- More realistic attitudes about the limitations as well as the potential contributions of such inquiry procedures as experimental design, standardized testing, surveys, and site visits
- Annual menus of continuing education opportunities for evaluators (especially those of the American Evaluation Association)
- Increasing emphasis on evaluation internationally and across disciplines

As seen in this book's previous chapters, the evaluation field has developed and vetted a number of alternative evaluation approaches, including Patton's utilization-focused evaluation, Scriven's Key Evaluation Checklist, Stake's responsive evaluation, and Stufflebeam's CIPP model for evaluation. Whereas Guba and Stufflebeam (1970) recommended the CIPP model as a basis for institutionalizing sound evaluation, in this chapter we recommend that the subject organization engage its staff to consider a range of potentially sound evaluation models and then choose and adopt (and, as appropriate, adapt) the one that best fits the organization's mission and culture.

Clearly, today's organizations can draw on a much richer supply of evaluation experiences, ideas, approaches, personnel, tools, methods, and training opportunities than was available in the early 1970s.

Checklist for Use in Institutionalizing and Mainstreaming Evaluation

This section presents and explains a checklist designed to guide organizations through a process of planning and installing a sound organizational evaluation system. The checklist includes fifteen checkpoints for consideration in establishing an organizational evaluation system that is fully functional. It is reflective of the first author's present collaboration with Michael Coplen to help the Office of Research and Development (R&D) of the Federal Railroad Administration (FRA) develop, install, and employ an evaluation system for use in assessing and strengthening safety throughout the U.S. railroad industry. In their development and use of the checklist, Coplen and Stufflebeam are employing a participatory approach. They are being aided by the Office of R&D's director (a consummate evaluation-oriented leader); by a review panel of the chiefs of the office's four divisions (in our terms, evaluation-oriented researchers); and by a stakeholder evaluation review panel whose membership is reflective of groups within government (especially the U.S. Department of Transportation's Safety Council) and the railroad industry (including labor and management). We view such stakeholder involvement as absolutely crucial in ensuring the success of efforts to institutionalize and mainstream evaluation.

The checklist provided in this section is generic, and not locked in stone. We advise organizations to use it as a general guide, not a fixed set of prescriptions. The point needs to be underscored that an organization's administration should not "lay on" a common approach to evaluation. Administrators should work out their approach with representatives of all the stakeholders (in other words, those who will implement or use results from the evaluation system). The checklist is offered as a tool to help guide an organization's stakeholders in their deliberations to design and install a unified evaluation system.

Exhibit 26.1 contains the Checklist for Institutionalizing and Mainstreaming Evaluation. Its fifteen checkpoints are subsequently explained.

Exhibit 26.1 CHECKLIST FOR INSTITUTIONALIZING AND MAINSTREAMING EVALUATION

_____ 1. Establish evaluation system design and review teams.

_____ 2. Ground the evaluation system in pertinent professional standards for evaluations.

_____ 3. Adopt and define an evaluation approach that the organization's leaders and staff understand, value, and find useful.

_____ 4. Provide an appropriate and sufficient budget for evaluations.

_____ 5. Staff the evaluation function, including evaluation-oriented leaders, line- and staff-level personnel, evaluation specialists, external evaluators, and metaevaluators.

_____ 6. Conduct pilot tests of the evaluation system to strengthen its structure and enhance acceptance by stakeholders.

_____ 7. Include an overview of the planned evaluation system in official organizational documents, such as the organization's strategic plan and organizational brochures.

_____ 8. Prepare an official organizational evaluation system manual, and keep it current and responsive to the organization's evaluation needs.

_____ 9. Provide the organization's personnel with evaluation training and access to consultants as needed.

_____ 10. Promote service to the full range of pertinent evaluation users, both inside and outside the organization.

_____ 11. Assess all organizational components and factors that influence the organization's success, especially policies, the organization's strategic plan, budgets, programs, projects, management, and technical support.

_____ 12. Assess programs against the full range of proper evaluative criteria, such as responsiveness to client needs, excellence, viability, intended and unintended outcomes, cost-effectiveness, sustainability, and transportability.

_____ 13. Deliver feedback and formal reports for improvement as well as accountability purposes.

_____ 14. Continue to explain and "sell" the evaluation system to the organization's personnel and other stakeholders.

_____ 15. Subject the evaluation system to periodic internal and external reviews, and use the reviews to strengthen the system.

1. Establish Evaluation System Design and Review Teams

Organizations are advised to begin the process of developing an organizational evaluation system by appointing evaluation design and review teams. Both teams should be representative of the organization's personnel and constituents. Members could be appropriately drawn from the following: the organization's policy board, top and middle management, line- and staff-level personnel, technical and clerical staff, beneficiaries, and persons with evaluation experience and expertise.

The organization should define the evaluation design team's role basically as one of adapting, as appropriate, the Checklist for Institutionalizing and Mainstreaming Evaluation, then working through the checkpoints. The review team's role should be defined as that of a sounding board to provide members of the evaluation design team with reactions to their

draft materials; help them avoid tunnel vision; help them choose projects for pilot-testing the evaluation approach; and, especially, help ensure clarity, relevance, and feasibility of evaluation documents.

In possibly adapting the checklist, before working through it the evaluation design team should examine each checkpoint to determine whether it conforms to the organization's philosophy and makes practical sense considering the organization's mission and culture. Then they should revise, drop, replace, or expand the checkpoints as appropriate. In finalizing the checklist they should ensure that it is consistent with the organization's values and responsive to stakeholder evaluation review team feedback. Through this process, both teams ideally come to acquire a spirit of ownership of the process and its examined products.

2. Ground the Evaluation System in Pertinent Professional Standards for Evaluations

In beginning its work, the evaluation team is advised to determine the standards to be followed in conducting and assessing evaluations. Such standards provide the organization's personnel with common criteria for guiding and judging evaluations. An organization can use the standards to set evaluation policies, train staff, and give the public and external groups a basis for assessing the organization's evaluations. The organization's personnel will find that their evaluation efforts are greatly facilitated when they systematically follow the guidance contained in standards for evaluations. For assistance with choosing, adapting, and applying the organization's evaluation standards, the evaluation design team is advised to review Chapter 3 and use it as a convenient reference.

At present, the FRA, Office of R&D project to institutionalize and mainstream evaluation is being grounded in the program evaluation standards of the Joint Committee on Standards for Educational Evaluation (2011), which call for evaluations to meet requirements for utility, feasibility, propriety, accuracy, and evaluation accountability. As seen in Chapter 3, however, an evaluation capacity development project might also involve choosing, adopting, and applying a different set of standards for evaluations, such as the U.S. Government Accountability Office's *Government Auditing Standards* (2007) or the American Evaluation Association's *Guiding Principles for Evaluators* (2004). In determining standards for guiding and judging its evaluations, an organization might usefully combine elements from different sets of standards. For example, in adapting the Joint Committee standards, an organization might incorporate the Independence standard from the government auditing standards.

3. Adopt and Define an Evaluation Approach That the Organization's Leaders and Staff Understand, Value, and Find Useful

A functional evaluation approach will be grounded in a sound definition of evaluation and provide an easily understood, valid framework for planning and conducting evaluations.

The organization's personnel need to agree on a common definition of evaluation. In doing so they should avoid adopting any of the misleading or otherwise dysfunctional definitions of evaluation. For example, sound and fully functional evaluation is more than measurement,

more than judgment by an expert or group of experts, and more than determining whether goals have been achieved. Moreover, it is not the same as empowerment or public relations. Nor should evaluation be equated with any specific procedure, such as a case study, survey, or experimental design.

If the organization does not settle on a definition of evaluation, it is likely that different members will work from different definitions. The result will be confusion and a lack of a unified evaluation approach. Although the organization's personnel can choose from legitimate alternative definitions of evaluation, we recommend the following definition, which agrees in general with the one presented by the Joint Committee (1981, 1994, 2011):

Evaluation is the assessment of the merit and/or worth of a program or some other object.

This definition focuses on the root term in evaluation, which is *value*. It says that evaluations should assess the value dimensions of merit and worth. Merit concerns a program's excellence. Worth concerns its cost-effectiveness in meeting clients' needs. Ideally, a particular program, project, or service should have excellent potential to serve the organization's clients, and it should possess worth, being able to effectively address the clients' needs at a reasonable level of expense.

Just as it needs a sound definition of evaluation, an organization also should adopt a sound framework for evaluation. We recommend that the evaluation design team choose, and as needed adapt, one of the following approaches: the CIPP model, the consumer-oriented approach, the responsive evaluation approach, or the utilization-focused approach. These approaches are explained in Chapters 13 through 16, respectively. By way of example, the FRA, Office of R&D evaluation institutionalization and mainstreaming project referenced earlier is applying an adaptation of the CIPP model and selectively is employing components of context, input, implementation, and outcome evaluation. (Note that FRA's Office of R&D adapted the CIPP model by giving the model's process and product evaluation components labels of implementation evaluation and impact evaluation.)

4. Provide an Appropriate and Sufficient Budget for Evaluations

An organization should provide a sufficient budget to enable its evaluation system to strongly support decision making and accountability at all levels. The budget should cover evaluation training for staff as well as costs associated with conducting evaluations.

In our experience, organizations that have operated excellent evaluation systems have devoted on the order of 2 percent of their total organizational budget to evaluation. Also, we have observed that organizations that have devoted less than 1 percent of their organizational budget to evaluation typically have had a weak evaluation system. We offer these observations not as hard-and-fast, research-based findings, but as general, ballpark notions of how much money an organization needs to invest in its evaluation system. Our observations concerning budgeting for organizational evaluation systems are based on extensive field service in helping organizations set up or evaluate and strengthen evaluation systems. An organization should first carefully plan its evaluation system, then determine the level of resources that will be required to ensure the system's soundness and effectiveness. The ballpark figure of 2 percent may help the organization gauge the adequacy of its budgetary projections.

Moreover, in regard to conducting specific project evaluations, our ballpark recommendation—based on experiences in evaluating projects—is that on average about 7 percent of a project's budget should be spent on formative and summative evaluation. Again, this is not a hard-and-fast rule, but it may be helpful in budgetary deliberations in the course of planning specific project evaluations.

An organization, in meeting its evaluation system's financial needs, should apply a prudential criterion. All organizations have limited resources to invest in evaluation. They should carefully allocate these resources to focus on the most important evaluative feedback needs. They should collect information only when they will use it and do so as efficiently as possible. Organizations never can completely automate their evaluation system. However, they can and should take all possible steps to increase the efficiency and cost-effectiveness of their evaluations.

5. Staff the Evaluation Function, Including Evaluation-Oriented Leaders, Line- and Staff-Level Personnel, Evaluation Specialists, External Evaluators, and Metaevaluators

In any organization, the most important evaluators are the decision makers. These include, especially, top and middle management and line- and staff-level personnel. In making and executing decisions, an organization's leaders and other staff have to constantly evaluate—to both guide their work and meet accountability requirements. Accordingly, the organization needs to define these leaders' evaluation roles and provide them with needed evaluation training and support.

It is often feasible and warranted to employ one or more evaluation specialists, in addition to the organization's evaluation-oriented administrators and staff. An organization of small or modest size might engage only a single evaluation specialist. In such a case, the organization should provide the coordinator with needed technical and clerical support as well as authorization to hire outside evaluation consultants on an as-needed basis. A large organization, in contrast, might employ a whole team of evaluation specialists. Such a team could include a director; field data collection personnel; clerical personnel; and specialists in measurement, statistics, reporting, and computer technology. Also, in staffing the evaluation function it is important to plan for and fund external evaluation functions, especially for conducting periodic metaevaluations focused on assessing and strengthening interval evaluations.

The key point for the evaluation design team here is to determine the planned evaluation system's personnel requirements and propose recommendations on how the organization can best meet these requirements.

6. Conduct Pilot Tests of the Evaluation System to Strengthen Its Structure and Enhance Acceptance by Stakeholders

Once checkpoints 1 through 5 have been addressed, it is desirable to further engage stakeholders to help test and refine the evaluation approach. For this purpose the evaluation design team and

evaluation review team together should deliberate to select suitable projects for pilot-testing the evaluation approach. Such projects should be ones that the organization's leaders and staff see as needing formative and summative evaluations, that would provide an apt test of the evaluation system, and that can supply sufficient funds to support a pilot evaluation.

Desirably, this set of pilot evaluations should be representative of the organization's programs. To pilot-test the previously referenced FRA, Office of R&D effort to institutionalize and mainstream evaluation, a project has been selected from each division of the Office of R&D: Human Factors, Tracks and Grade Crossings, Rolling Stock, and Signals and Communications.

7. Include an Overview of the Planned Evaluation System in Official Organizational Documents, Such as the Organization's Strategic Plan and Organizational Brochures

The organization's administrators can exercise important leadership in gaining organization-wide understanding and acceptance of the planned evaluation system by according it the imprimatur of an official part of the organization. Including a brief, two- or three-page description of the planned evaluation system in the organization's strategic plan is an apt means of informing stakeholders about the intended uses of the evaluation system, the general evaluation approach, and the seriousness of the organization's commitment to institutionalizing and mainstreaming evaluation. Also, it might be beneficial to summarize the developing evaluation system in organizational brochures and funding proposals.

8. Prepare an Official Organizational Evaluation System Manual, and Keep It Current and Responsive to the Organization's Evaluation Needs

Following design, pilot testing, and adoption of its evaluation system, the organization should document the system in an official manual. This evaluation system manual should clearly explain the organization's rationale, standards, and model for evaluation. It should define staff responsibilities in regard to evaluation and describe ways and means of developing the organization's evaluation capacity. It should stress the importance of stakeholder involvement in both planning evaluations and using findings. It should include helpful sections on evaluation design, staffing, budgeting, and contracting. It should briefly address such technical topics as sampling, data collection, data management, and data analysis. It should emphasize the importance of evaluation use and describe effective ways of reporting evaluation findings. It should also stress the importance of subjecting evaluations to internal and external metaevaluations. It should project a process for periodically evaluating and improving the evaluation system. And it should describe the organization's approach to annual planning and funding of evaluations and overseeing and managing evaluation efforts. Preparation of this manual should be keyed to using it to provide the organization's personnel and external stakeholders with a clear orientation to the organization's commitment and approach to evaluation. The manual should be so constructed that all members of the organization will find it useful in planning, funding, conducting, and reporting on sound evaluations.

9. Provide the Organization's Personnel with Evaluation Training and Access to Consultants as Needed

The organization's leaders should help every staff member adopt an evaluation orientation. The organization should train its personnel to collect, analyze, and use data to guide decisions and prepare accountability reports. It should encourage them to think critically about the evaluation work and how they might constantly improve it. As feasible, the organization should also provide staff members with training, technical support, and useful evaluation materials. And the organization should recognize, celebrate, and reward outstanding conduct and uses of evaluation.

All of an organization's professional staff members require at least a basic level of orientation and training in the concepts and procedures of evaluation. Minimally, they should receive periodic training in the evaluation system's plan and manual. Other training may include internal evaluation workshops and staff seminars; national workshops, such as those sponsored by the American Evaluation Association; and on-the-job assistance by an evaluation adviser. Key evaluation topics include evaluation standards, evaluation models, logic models, contracting, design, budgeting, data collection, analysis, reporting, and metaevaluation, plus, especially, stakeholder involvement and uses of findings.

10. Promote Service to the Full Range of Pertinent Evaluation Users, Both Inside and Outside the Organization

The organization's evaluators should periodically clarify their intended audiences; obtain their inputs in regard to evaluation needs and intended uses of evaluation; and, within bounds of feasibility, design and deliver responsive evaluation services. Users of evaluation findings are both internal and external to the organization.

Internal evaluation audiences include the personnel who govern, manage, and implement the organization's programs and associated operations. The organization's internal evaluation system should assist these groups with such tasks as assessing institutional policies and goals, conducting needs assessments, building evaluation into strategic plans, writing evaluation plans for inclusion in funding proposals, designing and maintaining relevant databases, obtaining feedback of use in assessing and strengthening ongoing programs and projects, and preparing accountability reports.

External evaluation audiences include the organization's overseers and constituents. For example, evaluators in FRA need to identify and address the evaluation needs and requirements of such client groups as Congress, the U.S. Department of Transportation's Safety Council, Amtrak and other railroad companies, and railroad unions. All such groups would be interested in assessments of needs and problems in the railroad industry and of funded projects' impacts on railroad safety. A hospital's external audience for evaluation would include pertinent accrediting organizations, government and other oversight groups, and the public. The evaluation interests of such groups would vary widely but might include feedback from patients, the incidence of surgical mistakes, problems with interactions of drug prescriptions, control of infection, the qualifications of health professionals, adequacy of follow-up service to patients, sufficiency of equipment, efficient use of hospital space, management of patient records, and cost containment.

Although members of the audience will surely vary in regard to their interests in the organization, program, or service being evaluated and their related information requirements, they are likely to use evaluation reports in one of two main ways: for formative or summative purposes.

No public service organization, such as a police department, school district, postal service, or health service, should delete this checkpoint. The organization might clarify, expand, or otherwise strengthen the checkpoint. But the evaluation system's effectiveness and credibility will depend on how well the organization defines and addresses the evaluation needs of rightful audiences, both internal and external to the organization.

11. Assess All Organizational Components and Factors That Influence the Organization's Success, Especially Policies, the Organization's Strategic Plan, Budgets, Programs, Projects, Management, and Technical Support

Most organizations have a unique hierarchy of staff and work activities. The involved levels may include projects and programs, task groups (such as information technology teams and institutional self-study teams); departments and divisions; and the organization as a whole. As much as possible, the organization should foster and support reasonable amounts of evaluation at all organizational levels.

Some organizations evaluate their programs and services based mainly or only on results, such as student test scores, associate sales, or successful patient outcomes. This approach is too narrow, because it omits study of the organizational processes that contributed to the outcomes. An organization's evaluation system should look at every component and factor that influences organizational success. Among others, these are program goals, plans, and the planning process; operating programs and projects; finances; materials and equipment; uses of technology; staff participation in decision making; staff development; stakeholder involvement; the organizational calendar and work schedule; publicity and communication; leadership and supervision; internal evaluation; and organizational policies. Evaluations should also involve looking for a program's important positive and negative side effects.

An organization clearly does not need to evaluate everything every year. For example, it might reasonably evaluate its policies every three years. Nevertheless, over time the organization should evaluate and strengthen every aspect of the organization that has an impact on the extent and quality of its services.

12. Assess Programs Against the Full Range of Proper Evaluative Criteria, Such as Responsiveness to Client Needs, Excellence, Viability, Intended and Unintended Outcomes, Cost-Effectiveness, Sustainability, and Transportability

Organizations are advised to take an initially broad view of potentially applicable evaluative criteria, and then to select those that are most relevant to evaluating a given program or other evaluand. Example criteria are creativity in addressing problems, responsiveness to client needs, superiority and feasibility of planned interventions, correct implementation

of assignments, effective management, efficient use of resources, achievement of goals, cost-effectiveness of outcomes, freedom from negative side effects, sustainability and transportability of successful interventions, and credibility with constituent groups. In initially planning evaluations, evaluators should be divergent in considering a wide range of potential evaluative criteria. Subsequently, they should converge on the criteria that they and the client agree are most important in carrying out a given evaluation assignment.

13. Deliver Feedback and Formal Reports for Improvement as Well as Accountability Purposes

The evaluation system should provide for reporting evaluation findings effectively to help focus, plan, and guide programs, and should also provide for summative evaluations after a project or program has run its course. As much as possible, reporting of evaluation findings should be integrated as a naturally occurring part of the organization.

Formative evaluation reporting should be aligned with the information needs of program managers and staff. Managers and staff might require needs assessment reports to help in setting program goals and priorities, critiques of draft proposals, evaluation sections for funding proposals, briefing sheets assessing a program's progress, and reports of initial outcomes. Formative evaluation findings may be presented in a variety of ways—for example, via printed reports, PowerPoint slides, memorandums, telephone exchanges, in-person exchanges with decision makers, oral presentations, town hall meetings, and focus groups.

In addition to producing formative evaluation reports, an organization's evaluation system should provide its constituents with summative evaluation reports. Funding agencies, regulatory groups, accrediting organizations, taxpayers, and other constituents desire and have a right to receive formalized information on the organization's expenditures, the quality of services, and the extent and significance of program outcomes. An organization might meet its accountability requirements in several ways. One of the best ways is to provide evidence that the organization regularly evaluates all aspects of its operations and uses the information to guide improvement. Summative evaluation reports typically should include a formal full-length report, an executive summary, and a supporting technical appendix or separate technical report. Summative findings, like formative findings, may be delivered in a number of ways, using a variety of media.

Organizations should guard against issuing accountability reports that are inflated public relations devices. To be accountable, an organization must issue factual reports that show areas for development as well as strengths. In the long run, candid, honest reporting of problems as well as improvement efforts enhances an organization's credibility and encourages constituents to financially support or otherwise aid the organization's improvement process.

14. Continue to Explain and "Sell" the Evaluation System to the Organization's Personnel and Other Stakeholders

The organization's personnel should understand, accept, support, and apply the adopted evaluation approach if it is to function effectively. The evaluation system should be theoretically sound, well defined, sufficiently funded, validly instrumented, easy to explain and use, and

responsive to the organization's information needs. These attributes will matter little, however, if the organization's personnel are neither committed to making it work nor able to do so. It is essential to make sure that stakeholders understand the evaluation system; to ensure that they have had meaningful opportunities to help focus and shape it; and to sell them on the system's fairness, validity, feasibility, and utility.

Especially, policymakers, administrators, line- and staff-level personnel, clients, and other stakeholders should be engaged in defining, testing, and refining the evaluation system. The organization should help staff members, through training and mentoring, to learn about and make effective use of the evaluation system. The organization should also regularly and repeatedly use input from these groups to adopt appropriate evaluation policies and periodically review and improve the evaluation system. Clearly, it behooves the developers of an evaluation system to involve stakeholders in its creation, make sure their questions about the system are clearly addressed, engage them in the development process, provide them with needed financial and other support, and convince them that properly conducting and reporting on systematic evaluations are in their own and the organization's best interests. An evaluation's needs for financial support relate, for example, to personnel salaries, consultant stipends, evaluation instruments, office materials and other expenses, information technology hardware and software, evaluation library materials, and travel.

An organization may employ a number of concrete steps to meet needs associated with explaining and selling the evaluation system to stakeholders. Examples of such steps are as follows:

- Employ task groups to involve stakeholders in designing the organization's evaluation system and helping shape individual studies.
- Develop and disseminate a brochure on the system. The brochure's description of the system should help persons throughout the organization explain the organization's evaluation approach to others.
- Conduct public meetings and special meetings with key leadership groups to report evaluation findings and hear audience reactions.
- Appoint and meet regularly with a standing, representative accountability commission. The organization's leaders should consider engaging this commission to help set priorities for future evaluation studies, define evaluation questions of interest to sponsors and constituents, and propose criteria for evaluating particular organizational functions. The organization's leaders could also engage the commission to critique draft evaluation tools and reports, facilitate data collection, discuss findings, and help communicate the findings to the broader audience. FRA's Office of R&D employs such a commission in the form of its National Review Board, which comprises leading figures from government and the railroad industry.
- Identify and involve evaluation clients, including the organization's administrators and directors of projects and programs to be evaluated. Evaluators within the organization should engage these persons to help clarify the most important evaluation questions,

evaluative criteria, and audiences for reports. Doing so will help evaluators address questions of importance to audiences. The exchange will also prepare the client and other participants to understand, accept, and use the evaluation findings.

- Present symposia on the evaluation system at pertinent professional meetings as a means of obtaining critical feedback from evaluation and other experts. For example, FRA's Office of R&D presented such a symposium at the 2013 annual meeting of the American Evaluation Association.
- Periodically engage a communication expert to help improve the communication-related aspects of the evaluation system. The organization could ask the expert to assess and provide direction for strengthening all the communication aspects of evaluation discussed above.

15. Subject the Evaluation System to Periodic Internal and External Reviews, and Use the Reviews to Strengthen the System

A thesis of this book is that an organization should regularly evaluate and strengthen every organizational function that has an impact on the services clients receive. Evaluation is one of the most important of such functions. The organization should regularly evaluate its evaluation system and individual studies both to improve them and to demonstrate that evaluations are both sound and cost effective.

As discussed in Chapter 25, metaevaluations are critically important to the success and credibility of service organizations. They guard against evaluations that might mislead decision makers or gloss over serious problems. They are essential for instilling public confidence in an organization's evaluation reports. They are also needed to ensure that evaluations in which an organization invests are sufficiently helpful in improving services to warrant costs.

An organization can meet the metaevaluation requirement by carrying out steps such as the following:

- Regularly apply professional standards for evaluation and measurement to plan and assess the organization's evaluations.
- Encourage staff to conduct internal metaevaluations as a means of achieving quality assurance, and support them in doing so. Staff should learn from their evaluation experiences and attest to the soundness of their evaluations.
- Engage an external evaluator to advise staff on evaluation policies and evaluate the organization's evaluation plans and draft reports.
- Periodically engage an external metaevaluator to evaluate and provide recommendations for strengthening the organization's evaluations.
- Employ metaevaluation results to improve utility, propriety, feasibility, accuracy, and accountability of evaluations.

The checklist in Exhibit 26.1 and associated explanations throughout the preceding subsections are intended to define the requirements of a sound organizational evaluation

system. It is in the best interest of an organization that seeks to succeed in serving clients to carefully consider, and adapt as needed, each checkpoint. Each one is relevant to defining a sound vision for an organization's evaluation system.

Summary

We began this chapter by reminding readers of the book's nine themes. We then focused readers' attention on organizations' needs in regard to institutionalizing and mainstreaming systematic evaluation. Two fundamental definitions undergirding the remainder of the chapter were as follows:

- Institutionalization of evaluation refers to an organization's commitment to evaluation and actions to define, install, regularly operate, and use results from a relatively permanent evaluation system.
- Mainstreaming evaluation means that the evaluation system functions at all levels of the organization by assessing all facets that are vital to fulfilling the organization's mission, and that the organization's full range of personnel are engaged in the conduct and use of evaluations.

After referencing early work in the area of institutionalizing and mainstreaming evaluation, we offered practical advice for organizations to use in installing their own evaluation system. This advice was presented in fifteen checkpoints, all oriented to helping organizational leaders and their staff design, fund, staff, and apply an evaluation system that is grounded in standards of the evaluation field, practical, effective, and valued by the organization's stakeholders. In explaining the checkpoints, we shared a current effort to institutionalize and mainstream evaluation in the Federal Railroad Administration's Office of Research and Development.

Evaluation is everybody's business in an organization, and efforts to institutionalize and mainstream systematic evaluation should entail substantial stakeholder engagement. We hope administrators and staff in a wide range of organizations will find this chapter useful and practical for making evaluation a normal, beneficial component of their efforts to plan and deliver outstanding and respected services.

REVIEW QUESTIONS

1. What are key points in the rationale for institutionalizing and mainstreaming evaluation?
2. What principles should guide the development of an organizational evaluation system?
3. What model did Guba and Stufflebeam employ in their 1970 paper on institutionalizing evaluation, and how does their choice compare with this chapter's recommendation in regard to the use of evaluation approaches?

4. What are the definitions of institutionalizing evaluation and mainstreaming evaluation presented in this chapter? What distinguishes institutionalization of evaluation from mainstreaming of evaluation?
5. What are the responsibilities of the evaluation design team and the evaluation review team in a project to institutionalize and mainstream evaluation?
6. What standards should undergird an organization's development of an evaluation system, and what is this chapter's position on adopting versus adapting published sets of standards?
7. What are the key staffing needs in a sound organizational evaluation system?
8. What steps would you take to apply this chapter's checkpoint pertaining to "selling" an evaluation system?
9. What are rules of thumb to consider in funding an organization's evaluation system?

Group Exercises

Exercise 1

Suppose your group has been asked to make a presentation to an organization's administration and professional staff on the case for institutionalizing and mainstreaming evaluation. List up to ten key points for inclusion in the presentation.

Exercise 2

Suppose your group has been contacted to conduct a metaevaluation of an organization's evaluation system. List the most important questions you would address. List criteria that should be considered in judging such a system. Summarize the kinds of information your group would collect. Finally, outline in general terms the final metaevaluation report your group would present.

Exercise 3

Suppose your group has been asked to outline a unit on institutionalizing and mainstreaming evaluation that is to be included in a graduate course on evaluation for administrators. Assume that the course already has units on formative and summative evaluation, evaluation models, evaluation standards, evaluation procedures, reporting findings, and metaevaluation. What would your group recommend as unique topics for the unit on institutionalizing and mainstreaming evaluation?

Exercise 4

As a group, develop a generic outline for an organization's evaluation system manual.

Suggested Supplemental Readings

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Deming, W. E. (1986). *Out of the crisis*. Cambridge, MA: Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Finn, C. E., Stevens, F. I., Stufflebeam, D. L., & Walberg, H. J. (1997). A meta-evaluation. *International Journal of Educational Research, 27,* 159–174.

Guba, E. G., & Stufflebeam, D. L. (1970, June). *Strategies for the institutionalization of the CIPP evaluation model*. Keynote address given at the 11th Phi Delta Kappa Symposium on Educational Research, Columbus, OH.

Joint Committee on Standards for Educational Evaluation. (2011). *The program evaluation standards* (3rd ed.). Thousand Oaks, CA: Corwin Press.

Peters, T. J., & Waterman, R. H. (1982). *In search of excellence: Lessons from America's best-run companies*. New York, NY: Warner Books.

Sanders, J. R. (2002). Presidential address: On mainstreaming evaluation. *American Journal of Evaluation, 33,* 253–259.

Sanders, W. L., & Horn, S. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8,* 299–311.

Stufflebeam, D. L. (1983). The CIPP model for program evaluation. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 117–151). Norwell, MA: Kluwer.

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., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond, R. L., Merriman, H. O., & Provus, M. (1971). *Educational evaluation and decision making*. Itasca, IL: Peacock.

Stufflebeam, D. L., & Shinkfield, A. J. (1985). *Systematic evaluation: A self-instructional guide to theory and practice*. Norwell, MA: Kluwer.

Stufflebeam, D. L., & Webster, W. J. (1980). An analysis of alternative approaches to evaluation. *Educational Evaluation and Policy Analysis, 2*(3), 5–20.

Stufflebeam, D. L., & Webster, W. J. (1988). Evaluation as an administrative function. In N. Boyan (Ed.), *Handbook of research on educational administration* (pp. 569–601). White Plains, NY: Longman.

Tymms, P. (1995). *Setting up a national "value-added" system for primary education in England: Problems and possibilities*. Paper presented at the National Evaluation Institute, Kalamazoo, MI.

Webster, W. J. (1995). The connection between personnel evaluation and school evaluation. *Studies in Educational Evaluation, 21,* 227–254.

Webster, W. J., Mendro, R. L., & Almaguer, T. D. (1994). Effectiveness indices: A "value added" approach to measuring school effect. *Studies in Educational Evaluation, 20,* 115–155.

GLOSSARY

Accreditation A process administered by an accrediting association to examine the quality of an institution (or section of an institution) or institutional program against externally based, professional accrediting standards, and to decide on whether to certify that institution based on the level of quality. Typically the accreditation process includes the subject institution's self-study, a subsequent examination by a visiting panel of experts, and the accrediting association's eventual decision either to provide some level of accreditation for a given period of years or to deny accreditation.

Accuracy standards Standards for program evaluations requiring evaluators to ground conclusions on evidence that is sufficient to address the targeted questions, unbiased, valid, reliable, analyzed correctly, and reported in full; that (desirably) has been subjected to an independent audit; and about which the evaluator has been properly candid concerning limitations.

Accuracy standards of the U.S. Government Accountability Office Government reporting standards for performance audits stating that reports should provide evidence that is true, contain findings that are correctly portrayed, and be credible and reliable in all matters. A report should clearly indicate limitations of the data and contain no unwarranted conclusions or recommendations based on those data.

Active-reactive-adaptive process of evaluation Interactive discussion, advice, and general consultation that continues throughout a utilization-focused evaluation between the evaluator and the intended users. Michael Patton has stated that the utilization-focused evaluator must be active in purposefully identifying intended evaluation users and working with these users to formulate questions that will shape a study. The evaluator is reactive in focusing on the thinking of users and responding to their ideas. As the process unfolds, the evaluator responsively adapts aspects of the evaluation to accommodate situational dynamics and increased understanding by both evaluator and users.

Advance organizers Variables or dimensions used in evaluations to determine information requirements, organize findings, format reports, plan reporting sessions, and so on.

Advocate teams technique A technique for use in input evaluations whereby teams generate competing proposals for meeting targeted needs, and an independent team evaluates the alternative proposals against established criteria.

Alpha The magnitude of a Type I error for an H_A hypothesis. For example, if the level of statistical significance has been set at 0.01, then across a large number of replications of the study, 1 percent of them would be expected to reject H_A when it is true.

Alpha level of significance The upper bound on rejecting a hypothesis when it is true. For example, if the significance level is set at 0.05, then across a large number of replications of the study, 5 percent of them would be expected to reject H_A when it is true.

Amateur evaluation Evaluation by persons with minimal evaluation expertise but striving to conduct a credible study.

American Evaluation Association (AEA) An association formed in 1986 with the merging of the Evaluation Research Society and the Evaluation Network.

American National Standards Institute (ANSI) A U.S. organization recognized to accredit standards in approximately ten thousand service and product areas, including the standards for educational evaluations in North America released by the Joint Committee on Standards for Educational Evaluation.

Analysis of information A process of identifying and assessing the constituent elements of each set of obtained information and their relationships to clarify the information's dependability and meaning for answering particular questions.

Antecedents Background information, including any conditions existing prior to an evaluation that may be relevant to interpreting program outcomes.

Application of findings The use of evaluation findings. Although this step is under clients' control, evaluators should at least offer to assist in the application of findings.

Apportionment A process of allocating a finite set of resources to alternative evaluands in consideration of their relative merit and importance.

Assets Accessible expertise and services, usually in the local area, that could be used to help fulfill a program's purpose.

Audit documentation A U.S. government fieldwork standard for performance audits stating that auditors should prepare and maintain documentation related to planning, conducting, and reporting on an audit. This documentation should contain sufficient information to enable an experienced auditor who has had no previous connection with the audit to ascertain and assess the evidence being advanced to support the auditors' significant judgments and conclusions. Before issuing their report, auditors should gather available documentation and ensure it is sufficient to support the audit's findings, conclusions, and recommendations.

Audit objectives A U.S. government reporting standard for performance audits stating that the objectives of an audit should be clear, specific, neutral, measurable, and feasible.

Audit scope A U.S. government reporting standard for performance audits stating that the depth and coverage of the audit work should be sufficient to explain the relationship between the population of items sampled and the sample examined; identify organizations, geographical locations, and the period covered; and explain and document problems. Auditors also are expected to report significant constraints on the audit.

Beta The magnitude of a Type II error for a specified H_A (the alternative hypothesis). For example, if the beta level is set at 0.10, then across a large number of replications of the study, 10 percent of them would be expected to erroneously accept H_0 (the null or nil hypothesis) as true when H_A is true.

Bidders' conference A conference that provides all potential evaluation bidders with an equal opportunity to receive background information about the needed evaluation and to address questions to the sponsor's representatives.

Case study evaluation An in-depth depiction, description, and analysis of a particular program or other evaluand, either in progress or as it occurred in the past. The case study's focus is the case itself.

Case study evaluator responsibilities Obligations including bounding and conceptualizing the case; selecting a phenomenon, themes, or issues to study; seeking patterns of data to develop issues for study; triangulating key observations and bases for interpretation; selecting alternative interpretations to pursue; and developing assertions or generalizations about the case.

Catalytic authenticity An evaluation's fostering of an event or events in which stakeholders took appropriate action.

Checklists Lists of items, such as tasks or criteria, to consider when undertaking an evaluation; some aspect of an evaluation; or, more broadly, some other enterprise.

Clear A U.S. government reporting standard for performance audits requiring reports to be easy to read and understand and prepared in language that is as simple, straightforward, and nontechnical as the subject permits.

Collective case study A case study in which evaluators move further away from their particular case as they study a number of cases together, so that they can inquire into the phenomenon or population of interest.

Communication specialist A person with excellent communication and technical skills who ensures that evaluation reports are relevant to the interests of their target audiences and understood by them.

Competence A U.S. government auditing standard stating that the staff assigned to perform an audit or attestation engagement collectively should possess adequate professional qualifications for the tasks required. In more general terms, competence is a guiding principle requiring evaluators to provide stakeholders with skilled, effective services.

Concise A U.S. government reporting standard for performance audits requiring that reports not be longer than necessary to convey and support their message.

Congruence analysis Analysis whereby an evaluator employs observations and logical conclusions to examine discrepancies between expected and actual findings in the interest of determining whether what was intended occurred.

Connoisseurship and criticism An approach to evaluation, arising from methods used to assess artistic and literary works, that assumes that experts in a given area are able to offer capable analyses and vivid, penetrating reports that are not attainable from the more commonly applied evaluation approaches.

Constructivist evaluation An evaluation approach that rejects any aspects of objective truth, while emphasizing that a program's truth is constructed by collecting and analyzing perspectives of individuals who work within and for the program, plus those who use its services. Constructivist evaluators employ a subjectivist, relativistic epistemology and specifically reject the experimental design paradigm.

Consumer-oriented evaluation An approach whereby the evaluator is the enlightened surrogate consumer, drawing conclusions about the merit and worth of the program being assessed. The welfare of consumers is the focus of and justification for the evaluation, and evaluative conclusions are gauged against the consumers' assessed needs.

Consumer reports Independent assessments of the costs and relative merit of alternative objects of a given type that are available to consumers.

Context evaluation Evaluation involving assessment of needs, problems, assets, and opportunities to help decision makers establish defensible goals and priorities and help relevant users judge goals, priorities, and outcomes.

Context, input, process, and product (CIPP) model for evaluation A comprehensive approach to conducting formative and summative evaluations whereby evaluators, in conjunction with stakeholders, seek clear, unambiguous answers to pertinent questions about an enterprise's context, inputs, processes, and products. A central theme is to assess the extent to which a programmatic effort effectively serves targeted beneficiaries and does so within a framework of defined, appropriate values—and to assist in this process. Operationally, according to the CIPP model, evaluation involves delineating, obtaining, reporting, and applying descriptive and judgmental information about an object's merit, worth, significance, cost, safety, feasibility, and probity to guide decision making, support accountability, disseminate effective practices, and increase understanding of the involved phenomena.

Convincing A U.S. government reporting standard for performance audits requiring audit results to be responsive to the audit objectives and presented persuasively, with conclusions and recommendations following logically from the presented facts.

Cooperative evaluation agreement An arrangement whereby the evaluator and sponsor collaborate in conducting an evaluation.

Cost-benefit analysis As applied to program evaluation, a set of largely qualitative procedures used to discover a program's cost and determine returns on program investments in terms of economic objectives attained and broader social benefits perceived.

Cost-plus budget A formal evaluation agreement that includes the funds required to conduct an evaluation assignment, plus an additional agreed-on charge for the evaluator's services

that falls outside the sphere of the contracted evaluation. A cost-plus budget can be built into a grant, a fixed-price agreement, or a cost-reimbursable agreement. There are three types: (1) a cost-plus-a-fee budget, whereby the additional funds are used to help sustain the contracting organization; (2) a cost-plus-a-grant budget, whereby the additional funds are used to support program functions, funding graduate students, research on evaluation, or an evaluation conference, for example; and (3) a cost-plus-profit budget, typically employed by for-profit evaluation organizations to make financial gain from contracted evaluations.

Cost-reimbursable evaluation contract An agreement that the evaluator will account for, report, and be reimbursed for actual evaluation project expenditures.

Counterfactual What would have happened in the absence of a treatment or program. In experiments, the counterfactual is typically a control or comparison group.

Critical competitors Well-performing alternatives to an evaluand that might be less expensive or more effective.

Decision- and accountability-oriented studies Evaluations emphasizing that program assessments should be used proactively to help improve a program as well as retroactively to judge its value.

Defensible purpose A desired end that has been legitimately defined, and that is consistent with a guiding philosophy, set of professional standards, institutional mission, mandated curriculum, national constitution, or public referendum, for example.

Deliberative democratic evaluation A process that emphasizes a democratic, equitable, and principled approach to program evaluation. An evaluator using this approach stresses the importance of stakeholder engagement; strives to produce valid, reliable, and defensible information; and ultimately seeks defensible conclusions about a program's merit and worth.

Descriptive information The part of a final evaluation report that objectively describes the program's goals, plans, operations, and outcomes based on a reasonable array of credible and relevant sources.

Eclectic evaluation approaches Evaluation approaches in which an evaluator pragmatically draws from and selectively applies a wide range of methods and concepts to address the questions of designated stakeholders.

Empowerment evaluation An approach that gives stakeholders authority over the evidence, implementation, findings, conclusions, and reporting of an evaluation.

Empowerment under the guise of evaluation A practice whereby an external evaluator helps a client develop evaluation expertise, but allows the client to write, edit, and then release a report under the external evaluator's name.

Environmental analysis A process that involves gathering contextual information in the form of available documents and data concerning such matters as area economics, population characteristics, relevant projects and services, crime statistics, employment rates, political

dynamics, cultural and educational offerings, health care facilities, welfare programs, local governance, the area power structure, environmental safety, transportation services, utility services, bridges and other infrastructure, and the needs of the target population.

Equity An evaluand's affirmative and reasonable conformance to principles of justice, freedom, equal opportunity, and fairness for all involved personnel without imposing bias, favoritism, or undue hardship on anyone.

Ethical principles of evaluation Codes or standards of probity that govern the behavior of evaluators.

Evaluability assessment A type of assessment developed as a particular methodology for determining the feasibility of progressing with an evaluation of a program. Informal, qualitative data are collected to determine if the program would be served productively by proceeding with a systematic, quantitative evaluation, such as a comparative, randomized controlled field experiment.

Evaluand The object of an evaluation, especially a program, project, or organization.

Evaluation The systematic process of delineating, obtaining, reporting, and applying descriptive and judgmental information about some object's merit, worth, probity, feasibility, safety, significance, equity, sustainability, and/or transportability. The result of an evaluation process is usually a tangible product, especially a printed, summative evaluation report.

Evaluation accountability standards Evaluation standards requiring sufficient documentation of evaluations and that metaevaluations be conducted both internally (for improvement) and externally (for accountability).

Evaluation agreements Agreements that may be combined into a formal contract (applicable to external evaluations) or a less formal memorandum of agreement (better suited to internal evaluations). Agreements of both types should provide a framework of mutual understanding for proceeding with the evaluation work.

Evaluation approach A broad conceptualization of evaluation methodology and practice, encompassing multiple evaluation approaches or models.

Evaluation budget A detailed estimate of financial and associated resources required to implement the full range of proposed evaluation tasks within a given time period. This budget should convince the sponsor that the study is affordable and feasible, and that it will be performed at a high level of quality and professionalism.

Evaluation by pretext Evaluation that begins with the preferred conclusions of the evaluation sponsor and in which data are arranged to support predetermined outcomes.

Evaluation client A person or group that will use the results of an evaluation for some purpose, such as program selection, program improvement, or accountability to a sponsor. The client group includes the person who commissioned the evaluation as well as those who will attend to and use its results.

Evaluation constraints The practical constraints of many evaluation assignments that preclude meeting, or make it extremely difficult to fully meet, the standards associated with sound information collection or analysis.

Evaluation contract A legally enforceable written agreement between the evaluator and client concerning the evaluation's specifications, both parties' responsibilities for conducting the evaluation, the conditions and schedule for payments, and provisions for renegotiating the contract as appropriate.

Evaluation contracting A process in which evaluator and client establish a trusting relationship and formalize their agreements in a written statement that holds each party accountable and provides instructions for resolving disputes.

Evaluation coordinator A person who manages the work effort, usually of several evaluators or evaluation projects.

Evaluation design The specific set of procedures used in an evaluation, from which inferences about the evaluand will be made.

Evaluation designer A person who designs an evaluation, ensuring that it will address relevant questions, meet information requirements, and yield needed reports. This individual also lays out plans for staffing, housing, and funding the evaluation.

Evaluation grant A financial award to support a qualified evaluator in conducting a study that is of interest to the evaluator, contains societal value, lies within the sponsor's mission, and is seen as being fundable.

Evaluation ideologies alleged to be seriously flawed According to Michael Scriven, all of the following ideologies are flawed:

- *Managerial ideology*. Scriven's well-managed evaluation requires more than guidance from a competent administrator, because "self-serving, indulgent" managers and evaluators may impose rigid controls that would distort evaluation procedures, outcomes, and subsequent decision making.
- *Positivist ideology*. This ideology applies to evaluators who, in their attempts to remove bias from scientific works, overreact and portray science in general, and evaluation in particular, as value-free.
- *Relativistic ideology*. This ideology can be viewed as an overreaction to problems associated with the positivist ideology. Relativists hold that everything is relative, and there is no objective truth—a position that may deny the possibility of determining merit.
- *Separatist ideology*. This ideology is rooted in the denial or rejection of the proposition that evaluation is a self-referent activity, perhaps best reflected in the notion that evaluators should be totally independent of what is evaluated.

Evaluation request for proposal An organization's or sponsor's published or direct mail invitation to evaluators to submit a proposal to conduct a particular evaluation.

Evaluation research A special kind of applied research whose goal, unlike that of basic research, is not to discover knowledge but to test the application of knowledge.

Evaluation researcher A person who conducts research on evaluation toward the development and validation of sound evaluation theory. Such research may include testing the reliability, validity, and applicability of different evaluation methods; studying the costs of evaluations; studying salient cases where organizations have institutionalized and mainstreamed evaluation practices; examining practices in light of knowledge and principles from relevant disciplines; chronicling and examining the evaluation field's history; comparing evaluation practices in different cultures; considering work in its historical context; comparing and contrasting different evaluation degree programs; and examining the field's guiding assumptions and hypotheses.

Evaluation respondent A person who, within the scope of an evaluation, fills out forms, answers test questions, responds to interview questions, submits her or his work products, and/or allows her or his work to be observed.

Evaluation review panel A representative group from an organization and its environment whose functions include reviewing evaluation plans, assessing the progress of the study as it evolves, perusing and commenting on draft interim and final reports, and facilitating the collection of information and dissemination of findings.

Evaluation sponsor An individual, institution, or organization that initiates an evaluation and provides financial and other resources to ensure its satisfactory conduct.

Evaluation stakeholders Individuals or groups closely identified with a program and likely to be affected by changes arising from the evaluation.

Evaluation standard A principle commonly agreed to by experts in the conduct and use of evaluation, used to measure the value or quality of an evaluation.

Evaluation trainer A person who teaches evaluation to potential evaluators.

Evaluee A person who is an object of an evaluation.

Evidence A U.S. government fieldwork standard for performance audits requiring that sufficient, appropriate evidence be obtained to provide a reasonable basis for auditors' findings and conclusions.

Experimental studies Studies designed to determine the effects of a program or other planned intervention. Evaluators employ random assignment or matching procedures to assign beneficiaries or organizations to experimental or control groups, administer a treatment to the experimental group, contrast outcomes for the involved groups, and make inferences about the intervention's effects.

External impairments A subset of the U.S. government Independence auditing standard requiring audit organizations to identify possible external impairments and ways of addressing them via internal policies and procedures for reporting and resolving external impairments.

Feasibility standards Evaluation standards requiring that evaluation procedures be efficient, politically viable, relatively easy to implement, adequately funded, and cost effective.

Feedback workshop technique A method for systematically conveying draft interim findings to a program's leaders and staff (and possibly other designated stakeholders), guiding their discussion of the findings, obtaining their critical reactions to draft reports and other materials, supporting their use of findings, and using their feedback to update or strengthen evaluation plans and materials.

Fieldwork standards for performance audits U.S. government auditing standards pertaining to planning an audit; supervising staff; obtaining sufficient and relevant evidence; and preparing audit documentation.

Final synthesis According to Michael Scriven, a process that includes searching for appropriate decision rules; deriving prima facie criteria admissible in probative judgments; deriving criteria of goodness inherent in the classical definition of the evaluation object; assessing the needs and preferences of the client and beneficiaries; obtaining evidence of the object's status in regard to the criteria of merit, worth, and significance; weighting the criteria; profiling the results; deciding whether to try for a final synthesis; and, if warranted, combining the results to reach an overall conclusion.

Focus group technique A group interview approach developed by the consumer research field and used by evaluators predominantly to obtain and analyze the views of stakeholders concerning the merit and worth of a program or to obtain multiple perspectives on a given evaluation question. The function and expertise of the group moderator are central to the successful outcomes of the focus group meeting.

Form A U.S. government reporting standard for performance audits stating that auditors should prepare audit reports that communicate the results of each audit.

Formal evaluation An evaluation that is relevant, rigorous, designed and executed to control bias, consistent with appropriate professional standards, and otherwise made useful and defensible.

Formative evaluation Evaluation in which the evaluator assesses and assists with the formulation of program goals and priorities; provides direction for planning by assessing alternative courses of action and draft plans; and guides program management by assessing implementation and providing feedback on plans and interim results.

General theory of program evaluation A conceptual framework that covers a wide range of program evaluations; denotes their modal characteristics, including logic and processes of evaluative discourse; and describes in general how program evaluations should be assessed and justified.

Goal and roles of evaluation A concept expressing that although all evaluations have a unitary, unchanging goal (to determine value as objectively as possible), their roles may vary widely in the pursuit and clarification of constructive uses of evaluative data. Two prevalent roles involve

formative evaluation (to assist in developing and implementing a program) and summative evaluation (to assess a program's value once it has been developed).

Goal-free evaluation An evaluation in which the evaluator is kept ignorant of a program's goals so that he or she can uncover the full range of program outcomes regardless of what was intended. The goal-free evaluator searches for what actually is occurring in a program and for all of the program's effects, and examines processes and outcomes against consumers' assessed needs.

Government auditing standards A set of standards developed by the U.S. Government Accountability Office to help auditors assess and ensure the validity of reported information concerning the results of government-funded programs and the soundness of related systems of internal control.

Grading Judging an evaluand's merit by assigning it a grade, such as A, B, C, D, or F, or a rating, such as outstanding, excellent, good, fair, poor, or very poor.

Grounded theory A conceptual framework based on systematic, rigorous documentation and analysis of actual program evaluations and their particular circumstances.

Guiding principles for evaluators A set of five principles and twenty-five underlying normative statements adopted by the American Evaluation Association to guide evaluation practice.

Hypothetical principles Research-based principles for conducting program evaluations.

Improvement- and accountability-oriented approaches Approaches focused on determining a program's merit and worth. The functions of such approaches are to foster program improvement and accountability, help consumers make wise choices from among optional programs and services, and certify meritorious programs and institutions as suitable for use by consumers.

Independence A U.S. auditing standard requiring auditors and audit organizations to remain free of impairments to independence.

Informal evaluation An evaluation that is prone to being unsystematic, lacking in rigor, and possibly biased.

Information scope Requires evaluators to, ideally, collect information that has sufficient breadth to address the audience's most important questions while also supporting a judgment of merit and worth. Typically, evaluators should obtain information on all important variables.

Input evaluation Evaluation in which the evaluator assesses alternative program strategies, competing action and staffing plans, and associated budgets to determine their differential feasibility and potential cost-effectiveness in meeting targeted needs and achieving goals.

Institutionalizing evaluation A process whereby an organization defines, installs, regularly operates, and uses results from an evaluation system that is relatively permanent in the organization.

Instrumental case study A case study that provides insight into an issue needing resolution or a theory or procedure needing refinement.

Integrity/honesty An American Evaluation Association guiding principle for evaluators stating that evaluators should display honesty and integrity in their own behavior and attempt to ensure the honesty and integrity of the entire evaluation process.

Internal evaluation Work within organizations to address evaluation needs. Such work might involve assessing the organization's externally funded projects or assessing the merit and worth of organizational plans or operations.

Intrinsic case study A case study undertaken to give a better understanding of the inner nature of a particular case, irrespective of its possible extrinsic value.

Intrinsic evaluation Evaluation whereby the evaluator assesses a program's inner quality, regardless of its effects on beneficiaries, by examining such aspects as policies, goals, structure, facilities, equipment, plans, procedures, staff qualifications, and communication.

Judgmental information Evaluative conclusions based on a set of values or standards plus discussions of a program's strengths and weaknesses, and possibly including recommendations for improvement.

Key Evaluation Checklist An evaluation tool developed and continually refined by Michael Scriven for applying his evaluation approach. The checklist can be adapted for use in particular evaluations, including metaevaluations. Its rationale is that evaluation is essentially a data reduction process, whereby large amounts of data are obtained and assessed and then synthesized in an overall judgment of value or at least a profile keyed to selected checkpoints.

Mainstreaming evaluation A process whereby an organization's evaluation system comes to function at all levels of the organization, assessing all facets that are vital to fulfilling the organization's mission.

Management information systems Mechanisms used to supply managers with information needed to conduct and report on their programs. Political control may lead to the provision of information that aims to give political advantage.

Meta-analysis A form of quantitative synthesis of studies that addresses a common research question and yields a composite effect size across studies. In the program evaluation research context, this usually involves synthesis of findings from experiments that contrast findings from treatment and control conditions.

Metaevaluation Evaluation of an evaluation that helps the original evaluator detect and address problems; helps ensure the evaluation's quality; and, ultimately, helps reveal the evaluation's strengths and limitations. It is the process of delineating, obtaining, and applying descriptive and judgmental information—about the utility, feasibility, propriety, accuracy, and accountability of an evaluation and about its systematic nature, competent execution, integrity and honesty,

respectfulness, and social responsibility—to guide the evaluation and report its strengths and weaknesses. Metaevaluation is a professional obligation of evaluators. Ideally, it involves both internal and external assessment of a given evaluation.

Metaevaluator A person who conducts metaevaluations. The role can be extended to involve assessment of the merit, worth, and probity of all that the evaluation profession is and does: evaluation services, evaluation tools, uses of evaluations, evaluation publications, evaluation training, evaluation research, and evaluation societies.

Methodology A U.S. government reporting standard for performance audits requiring auditors to clearly explain what was done to achieve audit objectives.

Methods-oriented studies Quasi-evaluations keyed to employing a particular method, such as experimental design, to evaluate a program.

Mixed-method approach An approach to program evaluation that involves employing a range of quantitative and qualitative methods.

Modular evaluation budgets Budgets that delineate the funding requirements for each part of a designed evaluation project or for each project year.

Naturalistic inquiry A procedure and process for studying a program as it unfolds, paying close attention to context, internal dynamics, and stakeholder insights. It imposes no controls on the development and delivery of the program or the assignment and involvement of participants. It involves investigatory categories and variables that evolve during the course of the study. The approach therefore minimizes investigator manipulation of the study setting and places no constraints on what the outcomes of the research will be. Typically, evaluators conducting naturalistic studies rely heavily on qualitative methods while also using quantitative techniques as deemed appropriate.

Need(s) Anything essential for a satisfactory mode of existence. It follows that anything without that condition would fall below a satisfactory level. Or, those things necessary or useful for fulfilling a defensible purpose.

Needs assessment A study to determine deficiencies in the well-being or performance of targeted beneficiaries or in the instrumentalities needed to prevent beneficiaries from suffering bad consequences. Determinations of such outcome and treatment needs provide a basis for setting goals and determining criteria for judging a program's outcomes.

Numerical weight and sum (NWS) A relatively common method for reaching evaluative conclusions, requiring computation of an overall score on an evaluation object by summing across all criteria the products of each criterion's weight times the object's score on the criterion. (This procedure could erroneously give a passing grade to an object that failed or did poorly on the most important criteria but scored high on less important or even trivial criteria.)

Objectives-based evaluation A classic example of quasi-evaluation in which the evaluator determines whether the program's objectives have been achieved. It is especially applicable in assessing tightly focused projects that have clear objectives.

Objective testing Testing that involves standardized procedures, such as the use of multiple-choice questions, to assess achievements of individual students or groups of students compared with norms, standards, or previous performance.

Objectivist evaluation Evaluation based on the theory that moral good is objective and independent of personal or simply human feelings. Fundamentally, objectivist evaluations are intended to lead to conclusions that are correct—not correct or incorrect relative to an evaluator's or other party's predilections, position, preferences, or point of view.

Objectivity A U.S. government reporting standard for performance audits stating that the entire report should be balanced in content and tone; shortcomings should be set in an appropriate context; and evidence should be presented in an unbiased, fair manner so that users can be persuaded by the facts.

Ontological authenticity An evaluation's success in helping stakeholders surface and understand their unconscious or unstated beliefs and values concerning a program.

Opportunities Advantageous circumstances, especially including funding programs that could be used to help fulfill targeted needs.

Organizational impairments A subset of the U.S. government Independence auditing standard stating that audit organizations need to be free from impairments to independence that might result from their place within or relationship to the organization that houses the entity to be audited.

Outcome need A level of achievement or outcome in a particular area required to fulfill a defensible purpose.

Pandering evaluations Studies that inform a client of what he or she wants to hear (often evading the truth of a program), toward the goal of winning the client's favor.

Payoff evaluation Evaluation concerned not with the structure or implementation of a program, but with its effects on beneficiaries (pertaining, for example, to test scores, employment, physical fitness, home ownership, or financial gains).

Personal impairments A subset of the U.S. government Independence auditing standard requiring audit organizations to maintain internal quality control systems to detect whether auditors have any relationships and beliefs that might cause them to be partial or give the appearance of partiality.

Personnel evaluation A systematic assessment of a person's qualifications or performance in relation to a role and defensible purpose of an institution, profession, program, or other entity.

Personnel evaluation standards A set of standards developed by the Joint Committee on Standards for Educational Evaluation and approved by the American National Standards Institute. The individual standards are grouped according to four essential attributes of sound personnel evaluation: utility, feasibility, propriety, and accuracy.

Phenomenology A concept pertaining to outward manifestations, that is, things perceptible by the senses, while emphasizing systematic scientific study of relationships between organizations (and their constituent individuals) and their environment.

Planning A U.S. government fieldwork standard for performance audits that directs auditors to define an audit's objectives, scope, and needed methods.

Politically controlled studies Studies in which the client seeks the truth in regard to a program but may inappropriately control the release of findings.

Posttest-only designs Experimental designs that randomly assign participants to alternative treatment conditions, followed by posttreatment assessment and interpretation of outcomes for the different groups. Random assignment of subjects to treatment groups is recommended, instead of employing matching of subjects, to help ensure equivalence between groups and support the drawing of defensible causal inferences.

Pragmatic principles Procedural guidelines for conducting evaluations that have been shown to work well in evaluation practice.

Preordinate evaluation Evaluation that is usually focused narrowly on examining the extent to which preestablished objectives are achieved.

Pretest-posttest designs Experimental designs in which participants are randomly assigned to treatment and control groups, and measures are taken before and after the treatment.

Prescriptive theory A conceptual framework proposed by an evaluator—based on reflections on and critical analyses of a wide range of evaluation experiences—intended to act as a guide for designing and conducting evaluations.

Probative inference A prima facie conclusion about a program's value based on close study of relevant facts and context.

Process evaluation Evaluation in which the evaluator assesses the implementation of program plans, first to help staff carry out activities and thereafter to help a broader range of users judge program implementation. Through documentation of processes and reporting on progress to appropriate program staff members and other interested parties, the evaluator makes a judgment about the extent to which planned activities are being (or were) carried out on schedule, as planned, and efficiently.

Product evaluation In the context, input, process, and product (CIPP) model, evaluation in which the evaluator identifies and assesses outcomes—intended and unintended, short term and long term—to help staff keep an enterprise focused on achieving agreed-on, important outcomes and ultimately to help relevant users gauge the success of the effort in meeting targeted needs. In assessments of consumer products, this term refers to evaluation of such tangible products as computers, automobiles, and cameras. In product evaluations of the latter type, it is important to identify and validate criteria of merit; weight them according to their relative importance; assess the product on each criterion; and, where possible, reach an overall conclusion.

Professional evaluation Evaluation undertaken by trained evaluators possessing high-level technical skills, knowledge of evaluation theory and methodology, and a commitment to meeting the evaluation field's standards.

Professional judgment A U.S. government auditing standard requiring auditors to assess situations or circumstances and draw sound conclusions to serve the public interest effectively, all the while maintaining utmost integrity, objectivity, and independence.

Program evaluation Systematic collection, analysis, and reporting of descriptive and judgmental information about the merit and worth of a program in terms of its goals, design, processes, and outcomes to address improvement, accountability, and dissemination questions and increase understanding of the involved phenomena.

Program evaluation model An evaluation theorist's idealized conceptualization, ideally based on extensive, real-world evaluation experience, for conducting program evaluations.

Program evaluation standards A set of standards developed by the Joint Committee on Standards for Educational Evaluation and approved by the American National Standards Institute. The thirty standards are grouped according to five essential attributes of a sound evaluation: utility, feasibility, propriety, accuracy, and evaluation accountability.

Project profiles Sets of information characterizing projects, including such items as a project's mission, constituents, needs, plans, resources, and accomplishments to date.

Propriety standards Evaluation standards requiring that evaluation be conducted legally, ethically, and with due regard for the welfare of the affected parties, including beneficiaries as well as service providers.

Pseudoevaluations Studies in which evaluators purposely produce and report invalid assessments or withhold or selectively release findings to right-to-know audiences. In these instances, evaluators do not adhere to professional standards for evaluation. Six predominant types of pseudoevaluations are public relations studies, politically controlled studies, pandering evaluations, evaluation by pretext, empowerment under the guise of evaluation, and customer feedback evaluation.

Public relations studies Studies in which the emphasis is placed not on truth seeking, but on the acquisition and broadcasting of information that provides a favorable, but often spurious, impression of a program.

Purposive sampling Sampling that allows the evaluator to focus on key informants to obtain information. In such situations, random sampling would not be applicable. Purposive sampling may lead to snowball sampling, however, whereby initial interviewees identify others who could provide relevant information. It is good practice for investigators to apply both purposive sampling and random sampling, and then to compare the two sets of findings.

Qualitative analysis The process of compiling, analyzing, and interpreting qualitative information to answer particular questions about a program.

Quality Control and Assurance A U.S. government auditing standard stating that each organization that performs audits or attestation engagements should have an appropriate internal system in place for maintaining a high degree of excellence, and should undergo periodic external peer reviews.

Quantitative analysis Analysis involving a wide range of concepts and techniques for using quantitative information to describe a program and to study and communicate its effects. The process involves compiling, exploring, validating, organizing, summarizing, analyzing, synthesizing, and interpreting quantitative information. Quantitative analysis techniques include descriptive, inferential, and nonparametric statistical techniques, among many others.

Quasi-evaluations Legitimate evaluation studies that sometimes are too narrow in regard to the questions addressed or methods employed to support an assessment or conclusion concerning a program's merit and worth.

Quasi-experimental designs Designs employed when an evaluator is interested in exploring the causal relationships between a program and its outcomes but random assignment is not feasible. These designs include various safeguards to counter threats to internal validity.

Questions-oriented program evaluation A quasi-evaluation approach whereby the evaluator addresses specified questions, often employing a wide range of methods.

Random assignment A method used in experiments for assigning participants to a treatment or comparison group. In random assignment, each participant should have the same probability of being assigned to any given condition (for example, a treatment group or a control group).

Random sampling A type of probability sampling in which each sample of size n from the population of interest has an equal, known probability of being selected for data collection. This method maximizes the likelihood that the sample will be representative of the population and thus permits generalization from the sample to the population. Random sampling allows the evaluator to set acceptable confidence intervals and draw a sample of sufficient size to achieve the desired level of precision for the estimate. If the requirements of random sampling are met (and valid measures are obtained), the evaluator can reach defensible inferences about certain features of the target population.

Random selection A sampling procedure in which members of a population are randomly selected to participate in a study. In random selection, all samples of size n taken from a population of size N have the same probability of selection.

Ranking Placing different institutions, programs, or persons in an ordered list according to their scores on a criterion of merit.

Realist evaluation A concept of evaluation calling for sustained, long-term study of a particular social intervention—such as Head Start—to develop explanations of why, how, where, and for whom the approach works or fails to work.

Regression discontinuity design A quasi-experimental design whereby subjects who meet a certain criterion are compared with those who failed to meet that criterion. A difference in the regression line for the two groups suggests a program effect.

Reliability A condition that is achieved when obtained information is free from internal contradictions and when repeated information collection episodes yield, as expected, the same answers.

Report contents A U.S. government reporting standard for performance audits stating that the audit report should include the objectives, scope, and methodology of the audit; the audit results, including findings, conclusions, and recommendations, as appropriate; a reference to compliance with generally accepted government auditing standards; the views of responsible officials; and, if applicable, the nature of any privileged and confidential information omitted.

Report distribution A U.S. government reporting standard for performance audits stating that auditors should submit reports to the appropriate officials of the audited entity and the appropriate officials of the one or more organizations requiring or arranging for the audit.

Report quality element A U.S. government reporting standard for performance audits stating that audit reports should be timely, complete, accurate, objective, convincing, clear, and as concise as the subject permits.

Reporting Effectively and accurately communicating an evaluation's findings in a timely manner to interested and right-to-know audiences.

Reporting findings A U.S. government reporting standard for performance audits stating that results should be keyed to the audit objectives and supported by sufficient, appropriate evidence.

Reporting standards for performance audits U.S. government auditing standards that pertain to the form of reports, report contents, report quality, and report issuance and distribution.

Respect for people An American Evaluation Association guiding principle for evaluators requiring evaluators to respect the security, dignity, and self-worth of respondents, program participants, the client, and other evaluation stakeholders.

Responsibilities for general and public welfare An American Evaluation Association guiding principle for evaluators requiring evaluators to articulate and take into account the diversity of general and public interests and values that may be related to an evaluation.

Responsive evaluation Also known as stakeholder-centered evaluation; a relativistic, social agenda and advocacy approach whereby the evaluator interacts with stakeholders (often a diverse group) to support and help develop, administer, and improve a program in a nondirective, counseling manner. An evaluator using this approach employs descriptive and judgmental information to examine a program's background, rationale, transactions, standards, and outcomes. Special features of this approach are searching for side effects, representing the inputs and judgments of diverse stakeholders, and issuing holistic reports.

Safety A concern leading to assessments of the risks associated with implementing and operating a program.

Scoring A process that involves assigning a number to a program or examinee. The number represents a sum of quality points that usually are assumed to be equal in value and additive.

Self-referent nature of evaluation The idea that as professionals, evaluators must evaluate and improve their services. This requires regularly evaluating their own work against professional standards and obtaining independent assessments of their evaluations.

Significance A concept referring to a program's potential influence, importance, and visibility.

Significance of evaluation The extent to which the evaluation field is contributing, in important ways, to society's welfare.

Single-object reports Reports that focus on a single program or other object. Typically, final reports of this kind are keyed to informing a broad audience about a program's background, structure, implementation, cost, main effects, and side effects.

Social agenda and advocacy approaches to evaluation Approaches in which evaluators direct their efforts toward increasing social justice through program evaluation, seeking to ensure that all segments of society have equal access to sound educational and social opportunities and services.

Social system An interrelated set of activities that ideally function together to fulfill a mission and achieve defined goals within a certain context.

Stakeholder-centered evaluation See *responsive evaluation*.

Stakeholders Those who are the intended users of an evaluation's findings, others who may be affected by the evaluation, and those expected to contribute to the evaluation. These persons are appropriately engaged in helping affirm foundational values, define evaluation questions, clarify evaluative criteria, contribute needed information, interpret findings, and assess evaluation reports.

Standards-based evaluation Evaluation that is grounded in and guided by professionally endorsed principles of sound evaluation.

Statistical hypothesis Typically, a statement about a population that one seeks to affirm or reject based on data from a sample of the population.

Statistical test An application of statistical procedures for deciding whether to accept or reject a hypothesis.

Statutory protections concerning auditing U.S. government–defined safeguards that protect against abolishment of the audit organization by an audited entity; require transparency of reasons for removing the head of the audit organization; prevent the audited entity from interfering in the audit; require the audit organization to report to a governing body that is independent from the audited entity; give the audit organization sole authority over staffing

the audit work; and guarantee the audit organization access to records and documents needed to complete an audit.

Structured observations Information gained when investigators systematically focus on and record observations of specific behaviors or characteristics of program personnel.

Success Case Method A quasi-evaluation approach based on emphasizing illuminated instances of program success, which are contrasted with the program's failing or failed elements.

Summative evaluation Evaluation that helps consumers decide whether a product or service—refined through development and formative evaluation—is a better buy than other alternatives. In general terms, summative evaluation typically occurs following the development of a product, completion of a program, or end of a service cycle. The evaluator draws together and supplements previous information and provides an overall judgment of the evaluand's value.

Supervision A U.S. government fieldwork standard for performance audits stating that audit staff are to be properly supervised.

Synthesis of information A process in which one or more methods are used to combine analysis findings across information collection procedures and devices to discern their validity and aggregate meaning for answering an audience's questions and judging the subject program's value.

Systematic evaluation Standards-based evaluation that is conducted with great care, with the evaluator not only (1) collecting information of high quality but also (2) clarifying and providing a defensible rationale for the value perspective used to interpret the findings and reach judgments as well as (3) communicating evaluation findings accurately to the client and other audiences.

Systematic information control An information management process to ensure that an evaluation's information is regularly and carefully checked, made as error-free as possible, and kept secure.

Systematic inquiry An American Evaluation Association guiding principle stating that evaluators should conduct systematic, data-based inquiries.

Tactical authenticity A study's success in advocating for all stakeholders, especially those with little power.

Technical specialist A person with technical expertise whose work supports the evaluation effort. This individual could be a test development specialist, sampling specialist, computer specialist, statistician, case study specialist, or technical writer. More than one technical specialist could be involved in a given evaluation.

Telephone interviews Interviews typically conducted by multiple interviewers within a phone bank who code responses as they are received. The interview protocol needs to be scripted so that all interviewers will obtain comparable data that can be aggregated and analyzed.

Theory-based evaluation Evaluation that begins with a well-developed and validated theory of how a defined program operates to produce outcomes or with an approximation of such a theory at an initial stage of a particular program evaluation.

Theory of evaluation A coherent set of conceptual, hypothetical, pragmatic, and ethical principles that form a general framework to guide the study and practice of evaluation.

Transdiscipline A discipline comprising a core field as well as a number of independent applied fields. Its principal mission is developing procedures and tools for use by a wide range of applied fields and disciplines. Statistics, measurement, logic, and evaluation are some of the more important examples.

Transdiscipline of evaluation A conceptualization of the evaluation field as encompassing evaluations of various entities across all applied areas and disciplines. The transdiscipline of evaluation comprises a common logic, theory, and methodology that transcend specific evaluation domains but that also have unique characteristics.

Transformative evaluation An evaluation approach that emphasizes social justice by giving precedence to the voices of the least advantaged groups in society.

Traveling observer technique A procedure developed by the Evaluation Center at Western Michigan University that directly addresses process evaluation data requirements while also yielding information that is useful in context, input, and product evaluations. A preprogrammed investigator, working on-site, investigates and characterizes how staff members are carrying out a project. Subsequently, this observer reports findings to other evaluation team members and assists them in planning follow-up site visits.

Treatment need A certain service, technique, tool, service provider, or other helping agent required to meet an outcome need.

Triangulation A process of reaching conclusions about the consistency of outcomes from varying sources and methods used for measuring a particular construct.

Type I error The probability of rejecting a given hypothesis (for example, the hypothesis that the difference between assessed outcomes of alternative treatments is zero) when the hypothesis is in fact true.

Type II error The probability of not rejecting an H_0 or null hypothesis when a given alternative hypothesis is true.

Unstructured observations Information obtained from loosely controlled, unobtrusive surveillance of program operations designed to help focus and structure later, more systematic observations.

Utility standards A set of evaluation standards stating that an evaluation should serve the information needs of its intended users.

Utilization-focused evaluation A form of eclectic evaluation developed by Michael Patton that is geared toward ensuring that evaluations have an impact. The utilization-focused evaluator

guides the evaluation process in collaboration with an identified group of priority users, placing focus squarely on their intended uses of the evaluation.

Validation The process of compiling evidence that supports the interpretations concerning and uses of data and information that were collected using one or more instruments and procedures. Validity resides not in any instrument or procedure, but in its use in generating inferences and conclusions in a particular study.

Value A defensible guiding principle or ideal that should be used to determine an evaluand's standing. A value might be one out of a number of ideals held by a society, group, or individual. As the root term in evaluation, *value* is central to determining the criteria for use in judging programs or other entities.

Value-added assessment A form of outcome evaluation depending on systematic assessment coupled with hierarchical gain score analysis to assess the effects of programs and policies. Emphasis is often on annual testing of students at various grade levels to assess trends and partial out the effects of different components of an education system (for example, individual schools and groups of schools).

Worth A program's combination of excellence, service, and cost-effectiveness in an area of clear need, within a specified context.

REFERENCES

Abma, T. A. (2006). The practice and politics of responsive evaluation. *American Journal of Evaluation*, *27*, 31–43.

Adams, J. A. (1971). *A study of the status, scope and nature of educational evaluation in Michigan's K–12 school districts*. Unpublished doctoral dissertation, Ohio State University, Columbus.

Alexander, D. (1974). *Handbook for traveling observers*. Kalamazoo: Western Michigan University, Evaluation Center.

Alkin, M. C. (1969). Evaluation theory development. *Evaluation Comment*, *2*, 2–7.

Alkin, M. C. (1985). *A guide for evaluation decision makers*. Thousand Oaks, CA: Sage.

Alkin, M. C. (1995, November). *Lessons learned about evaluation use*. Paper presented at the annual meeting of the American Evaluation Association, Vancouver, British Columbia, Canada.

Alkin, M. C. (Ed.). (2004). *Evaluation roots: Tracing theorists' views and influences*. Thousand Oaks, CA: Sage.

Alkin, M. C. (2011). *Evaluation essentials: From A to Z*. New York, NY: Guilford Press.

Alkin, M. C. (Ed.). (2013). *Evaluation roots: A wider perspective of theorists' views and influences* (2nd ed.). Thousand Oaks, CA: Sage.

Alkin, M. C., & Christie, C. A. (2004). An evaluation theory tree. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 12–66). Thousand Oaks, CA: Sage.

Alkin, M. C., Daillak, R., & White, P. (1979). *Using evaluations: Does evaluation make a difference?* Sage Library of Social Research, Vol. 76. Thousand Oaks, CA: Sage.

Alkin, M.C., & Taut, S.M. (2003). Unbundling evaluation use. *Studies in Educational Evaluation*, *29*, 1–12.

Altschuld, J. W., & Witkin, B. R. (2000). *From needs assessment to action: Transforming needs into solution strategies*. Thousand Oaks, CA: Sage.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Evaluation Association. (2004). *Guiding principles for evaluators*. Washington, DC: Author. Retrieved from http://www.archive.eval.org/Publications/GuidingPrinciples.asp

American Evaluation Association Task Force on High Stakes Testing. (2002). *Position statement on high stakes testing in pre-K–12 education*. Louisville, KY: Author.

Bamberger, M., Rugh, J., & Mabry, L. (2012). *RealWorld evaluation: Working under budget, time, data, and political constraints* (2nd ed.). Thousand Oaks, CA: Sage.

Bandura, A. (1977). *Social learning theory*. Upper Saddle River, NJ: Prentice Hall.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Barrett, P. (2011, February). *The use of sensory and analytical evaluation to match the flavor of onion powder in a white sauce*. Paper presented at the Evaluation Center's Evaluation Café, Western Michigan University, Kalamazoo.

Becker, M. H. (1974). The health belief model and personal health behavior. *Health Education Monographs*, *2*, 324–473.

Berkley, T., Day, G., Smith, S., & Chianca, T. K. (2005, October). *Using Success Case Method to assess hard to measures outcomes within a foundation-wide organizational development initiative*. Paper presented at the meeting of the American Evaluation Association and the Canadian Evaluation Society, Toronto, Ontario, Canada.

Beywl, W. (2000). Standards for evaluation: On the way to guiding principles in German evaluation. In C. Russon (Ed.), *The program evaluation standards in international settings* (pp. 60–67). Kalamazoo: Western Michigan University, Evaluation Center.

Bhola, H. S. (1998). Program evaluation for program renewal: A study of the National Literacy Program in Namibia (NLPN). *Studies in Educational Evaluation*, *24*, 303–330.

Bickman L. (Ed.). (1987). *Using program theory in evaluation*. New Directions for Program Evaluation, no. 33. San Francisco, CA: Jossey-Bass.

Bickman, L. (1996). The application of program theory to the evaluation of a managed mental health care system. *Evaluation and Program Planning*, *19*, 111–119.

Bickman, L., & Peterson, K. A. (1990). Using program theory to describe and measure program quality. In L. Bickman (Ed.), *Advances in program theory* (pp. 61–72). New Directions for Program Evaluation, no. 47. San Francisco, CA: Jossey-Bass.

Bickman, L., & Rog, D. J. (2009). *Handbook of applied social research methods* (2nd ed.). Thousand Oaks, CA: Sage.

Bigman, S. K. (1961). Evaluating the effectiveness of religious programs. *Review of Religious Research*, *2*, 108–109.

Birckmayer, J. D., & Weiss, C. H. (2000). Theory-based evaluation practice: What do we learn? *Evaluation Review*, *24*, 407–431.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I; Cognitive domain*. New York, NY: McKay.

Borenstein, M., Hedges, L. V., Higgins, J.P.T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, West Sussex, UK: Wiley.

Boruch, R. F. (1994). The future of controlled randomized experiments: A briefing. *Evaluation Practice*, *15*, 265–274.

Boruch, R. F. (2003). Randomized field trials in education. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 107–124). Norwell, MA: Kluwer.

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.

Brannick, M. T., & Levine, E. L. (2002). *Job analysis: Methods, research, and applications for human resource management in the new millennium*. Thousand Oaks, CA: Sage.

Brannick, M. T., Levine, E. L., & Morgeson, F. P. (2007). *Job and work analysis: Methods, research, and applications for human resource management* (2nd ed.). Thousand Oaks, CA: Sage.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

Brickell, M. (1976). *Needed: Instruments as good as our eyes* (Occasional Paper Series, Paper #7). Kalamazoo: Western Michigan University, Evaluation Center.

Brickell, M. (2011). Needed: Instruments as good as our eyes. *Journal of MultiDisciplinary Evaluation*, *7*(15), 171–179.

Brinkerhoff, R. O. (2003). *The Success Case Method: Find out quickly what's working and what's not*. San Francisco, CA: Berrett-Koehler.

Brinkerhoff, R. O. (2005a, October). *Making a causal argument for training impact*. Paper presented at the Evaluation Center's Evaluation Café, Western Michigan University, Kalamazoo.

Brinkerhoff, R. O. (2005b). The Success Case Method: A strategic evaluation approach to increasing the value and effect of training. *Advances in Developing Human Resources*, *7*, 86–101.

Brinkerhoff, R. O. (2006). *Telling training's story: Evaluation made simple, credible, and effective*. San Francisco, CA: Berrett-Koehler.

Brinkerhoff, R. O., & Dressler, D. (2002). Using evaluation to build organizational performance and learning capability: A strategy and a method. *Performance Improvement*, *41*(6), 14–21.

Brisolara, S. (1998). The history of participatory evaluation and current debates in the field. In S. Brisolara (Ed.), *Understanding and practicing participatory evaluation* (pp. 25–41). New Directions for Evaluation, no. 80. San Francisco, CA: Jossey-Bass.

Broom, L. (1964). Introduction. In A. Kaplan, *The conduct of inquiry* (pp. xxi–xxiii). San Francisco, CA: Chandler.

Brunner, I., & Guzman, A. (1989). Participatory evaluation: A tool to assess projects and empower people. In R. F. Conner & M. Hendricks (Eds.), *International innovations in evaluation methodology* (pp. 9–18). New Directions for Program Evaluation, no. 42. San Francisco, CA: Jossey-Bass.

Bryk, A. S. (Ed.). (1983). *Stakeholder-based evaluation*. New Directions for Program Evaluation, no. 17. San Francisco, CA: Jossey-Bass.

Candoli, I. C., Cullen, K., & Stufflebeam, D. L. (1997). *Superintendent performance evaluation: Current practice and directions for improvement*. Norwell, MA: Kluwer.

Campbell, D. T. (1975). Degrees of freedom and the case study. *Comparative Political Studies*, *8*(2), 178–193.

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. In W.M.K. Trochim (Ed.), *Advances in quasi-experimental design and analysis* (pp. 67–77). New Directions for Program Evaluation, no. 31. San Francisco, CA: Jossey-Bass.

Campbell, D. T. (1988). The experimenting society. In E. S. Overman (Ed.), *Methodology and epistemology for social science: Selected papers* (pp. 315–333). Chicago, IL: University of Chicago Press.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on training* (pp. 171–246). Skokie, IL: Rand McNally.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Skokie, IL: Rand McNally.

Cassaro, D. A. (2005). Lesbian, gay, bisexual, and transgender issues in evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 226–228). Thousand Oaks, CA: Sage.

Chalmers, I. (2007). The lethal consequences of failing to make use of all relevant evidence about the effects of medical treatments: The need for systematic reviews. In P. Rothwell (Ed.), *Treating individuals: From randomized trials to personalized medicine* (pp. 37–58). London, UK: Elsevier.

Chambers, R. (1992). Rapid but relaxed and participatory rural appraisal: Towards applications in health and nutrition. In N. S. Scrimshaw & G. R. Gleason (Eds.), *Rapid assessment procedures: Qualitative methodologies for planning and evaluation of health related programmes* (pp. 295–305). Boston, MA: International Nutrition Foundation for Developing Countries.

Chambers, R. (1994). The origins and practice of participatory rural appraisal. *World Development*, *22*, 953–969.

Chelimsky, E. (1985). Comparing and contrasting auditing and evaluation: Some notes on their relationship. *Evaluation Review*, *9*, 483–508.

Chelimsky, E. (1987). What have we learned about the politics of evaluation? *Evaluation Practice*, *8*, 5–21.

Chelimsky, E. (1997). The political environment of evaluation and what it means for the development of the field. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook* (pp. 53–68). Thousand Oaks, CA: Sage.

Chelimsky, E. (1998). The role of experience in formulating theories of evaluation practice. *American Journal of Evaluation*, *19*, 35–55.

Chen, H. T. (Ed.). (1989). The theory-driven perspective [Special issue]. *Evaluation and Program Planning*, *12*(4).

Chen, H. T. (1990). *Theory-driven evaluations*. Thousand Oaks, CA: Sage.

Chen, H. T. (1994). Theory-driven evaluation: Needs, difficulties, and options. *Evaluation Practice*, *15*, 79–82.

Chen, H. T. (1996). A comprehensive typology for program evaluation. *Evaluation Practice*, *17*, 121–130.

Chen, H. T. (2005). *Practical program evaluation: Assessing and improving planning, implementation, and effectiveness*. Thousand Oaks, CA: Sage.

Chen, H. T., & Rossi, P. H. (1983). Evaluating with sense: The theory-driven approach. *Evaluation Review*, *7*, 283–302.

Chen, H. T., & Rossi, P. H. (1987). The theory-driven approach to validity. *Evaluation and Program Planning*, *10*, 95–103.

Chen, H. T., & Rossi, P. H. (Eds.). (1992). *Using theory to improve program and policy evaluations*. Santa Barbara, CA: Greenwood Press.

Chianca, T. K., & Risley, J. S. (2005, October). *Applying the Success Case Method as part of the institutional evaluation of a nonprofit organization*. Paper presented at the meeting of the American Evaluation Association and the Canadian Evaluation Society, Toronto, Ontario, Canada.

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* (pp. 7–35). New Directions for Evaluation, no. 97. San Francisco, CA: Jossey-Bass.

Christie, C. A. (2011). Advancing empirical scholarship to further develop evaluation theory and practice. *Canadian Journal of Program Evaluation*, *26*, 1–18.

Christie, C. A., & Azzam, T. (2004). What's all the talk about? Examining EVALTALK, an evaluation listserv. *American Journal of Evaluation*, *25*, 219–234.

Christie, C. A., & Fleischer, D. N. (2010). Insight into evaluation practice: A content analysis of designs and methods used in evaluation studies published in North American evaluation-focused journals. *American Journal of Evaluation*, *31*, 326–346.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Clancy, T., with Horner, C. (1999). *Every man a tiger: The Gulf War air campaign*. Commander Series. New York, NY: Putnam.

Clark, D. L., & Guba, E. G. (1965, October). *An examination of potential change roles in education*. Paper presented at the Seminar on Innovation in Planning School Curricula, Warrenton, VA.

Clements, P., Chianca, T., & Sasaki, R. (2008). Reducing world poverty by improving evaluation of development aid. *American Journal of Evaluation*, *29*, 195–214.

Cochran, W. G. (1970). *Sampling techniques*. Hoboken, NJ: Wiley.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

Cohen, J. (1980). Some historical remarks on the Baconian conception of probability. *Journal of the History of Ideas*, *41*, 219–231.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Cook, D. L. (1966). *Program evaluation and review technique: Applications in education*. Washington, DC: U.S. Government Printing Office.

Cook, T. D. (2007). "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, *142*, 636–654.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Skokie, IL: Rand McNally.

Cook, T. D., & Gruder, C. L. (1978). Metaevaluation research. *Evaluation Quarterly*, *2*, 5–51.

Cook, T. D., & Reichardt, C. S. (Eds.). (1979). *Qualitative and quantitative methods in evaluation research*. Thousand Oaks, CA: Sage.

Cook, T. D., Scriven, M., Coryn, C.L.S., & Evergreen, S.D.H. (2010). Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven. *American Journal of Evaluation*, *31*, 105–117.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*, 724–750.

Cooksy, L. J., & Caracelli, V. J. (2005). Quality, context, and use: Issues in achieving the goals of metaevaluation. *American Journal of Evaluation*, *26*, 31–42.

Cooksy, L. J., & Caracelli, V. J. (2009). Metaevaluation in practice: Selection and application of criteria. *Journal of MultiDisciplinary Evaluation*, *6*(11), 1–15.

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Applied Social Research Methods Series, Vol. 2. Thousand Oaks, CA: Sage.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.

Cooperrider, D. L., & Whitney, D. (2005). *Appreciative inquiry: A positive revolution in change*. San Francisco, CA: Berrett-Koehler.

Cordray, D. S., & Pion, G. M. (2006). Treatment strength and integrity: Models and methods. In R. R. Bootzin & P. E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 103–124). Washington, DC: American Psychological Association.

Coryn, C.L.S. (2006). A conceptual framework for making evaluation support meaningful, useful, and valuable. *Evaluation Journal of Australasia*, *6*(1), 45–51.

Coryn, C.L.S. (2007a). *Evaluation of researchers and their research: Toward making the implicit explicit*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Coryn, C.L.S. (2007b). The holy trinity of methodological rigor: A skeptical view. *Journal of MultiDisciplinary Evaluation*, *4*(7), 26–31.

Coryn, C.L.S. (2009, September). *Contemporary trends and movements in evaluation: Evidence-based, participatory and empowerment, and theory-driven evaluation*. Paper presented at the Evaluation Center's Evaluation Café, Western Michigan University, Kalamazoo.

Coryn, C.L.S., Gugiu, P. C., Davidson, E. J., & Schröter, D. C. (2008). Assessing needs in hidden populations using respondent-driven sampling. *Evaluation Journal of Australasia*, *7*(2), 3–11.

Coryn, C.L.S., & Hattie, J. A. (2006). The transdisciplinary model of evaluation. *Journal of MultiDisciplinary Evaluation*, *3*(4), 107–114.

Coryn, C.L.S., Hattie, J. A., Scriven, M., & Hartmann, D. J. (2007). Models and mechanisms for evaluating government-funded research: An international comparison. *American Journal of Evaluation*, *28*, 437–457.

Coryn, C.L.S., & Hobson, K. A. (2011). Using nonequivalent dependent variables to reduce internal validity threats in quasi-experiments: Rationale, history, and examples from practice. In S. Mathison (Ed.), *Really new directions in evaluation: Young evaluators' perspectives* (pp. 31–39). New Directions for Evaluation, no. 131. San Francisco, CA: Jossey-Bass.

Coryn, C.L.S., Noakes, L. A., Westine, C. D., & Schröter, D. C. (2011). A systematic review of theory-driven evaluation practice from 1990 to 2009. *American Journal of Evaluation*, *32*, 199–226.

Coryn, C.L.S., Schröter, D. C., & Hanssen, C. E. (2009). Adding a time-series design element to the Success Case Method to improve methodological rigor: An application for nonprofit program evaluation. *American Journal of Evaluation*, *30*, 80–92.

Coryn, C.L.S., Schröter, D. C., Miron, G., Kana'iaupuni, S. K., Tibbets, K., & Watkins-Victorino, L. M. (2007). *A study of successful schools for Hawaiians: Identifying that which matters*. Kalamazoo: Western Michigan University, Evaluation Center.

Coryn, C.L.S., Schröter, D. C., Youker, B. W., & Bakerson, M. A. (2006). *Kalamazoo Public Schools Middle School Summer Enrichment Program: Year three evaluation*. Kalamazoo: Western Michigan University, Evaluation Center.

Coryn, C.L.S., & Scriven, M. (Eds.). (2008). *Reforming the evaluation of research*. New Directions for Evaluation, no. 118. San Francisco, CA: Jossey-Bass.

Coryn, C.L.S., Stufflebeam, D. L., Davidson, E. J., & Scriven, M. (2010). The Interdisciplinary Ph.D. in Evaluation: Reflections on its development and first seven years. *Journal of MultiDisciplinary Evaluation*, *6*(13), 118–129.

Coryn, C.L.S., Tarsilla, M., & Hobson, K. A. (2010, November). *The dependability of Campbell Collaboration, Cochrane Collaboration, and What Works Clearinghouse research reviews*. Paper presented at the annual meeting of the American Evaluation Association, San Antonio, TX.

Coryn, C.L.S., & Westine, C. D. (2013, November). *A decade of research on evaluation in review: What has been investigated, what has not been investigated, and what could be and should be investigated*. Paper presented at the annual meeting of the American Evaluation Association, Washington, DC.

Council on Foundations. (1993). *Evaluation for foundations: Concepts, cases, guidelines, and resources*. San Francisco, CA: Jossey-Bass.

Cousins, J. B. (1996). Consequences of researcher involvement in participatory evaluation. *Studies in Educational Evaluation*, *22*, 3–27.

Cousins, J. B. (2003). Utilization effects of participatory evaluation. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 245–265). Norwell, MA: Kluwer.

Cousins, J. B. (2004a). Commentary: Minimizing evaluation misuse as principled practice. *American Journal of Evaluation*, *25*, 391–397.

Cousins, J. B. (2004b). Crossing the bridge: Toward understanding use through systematic inquiry. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 319–330). Thousand Oaks, CA: Sage.

Cousins, J. B. (Ed.). (2007). *Process use in theory, research, and practice*. New Directions for Evaluation, no. 116. San Francisco, CA: Jossey-Bass.

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*, 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, UK: Falmer.

Cousins, J. B., & Shulha, L. M. (2008). Complexities in setting program standards in collaborative evaluation. In N. L. Smith & P. R. Brandon (Eds.), *Fundamental issues in evaluation* (pp. 139–176). New York, NY: Guilford Press.

Cousins, J. B., & Whitmore, E. (1998). Framing participatory evaluation. In E. Whitmore (Ed.), *Understanding and practicing participatory evaluation* (pp. 5–23). New Directions for Evaluation, no. 80. San Francisco, CA: Jossey-Bass.

Covert, R. W. (1995). A twenty-year veteran's reflections on the guiding principles for evaluators. In W. R. Shadish, D. L. Newman, M. A. Scheirer, & C. Wye (Eds.), *Guiding principles for evaluators* (pp. 35–45). New Directions for Program Evaluation, no. 66. San Francisco, CA: Jossey-Bass.

Crabtree, B. F., & Miller, W. L. (1992). *Doing qualitative research*. Thousand Oaks, CA: Sage.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Cronbach, L. J. (1963). Course improvement through evaluation. *Teachers College Record*, *64*, 672–683.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*, 116–127.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco, CA: Jossey-Bass.

Cronbach, L. J., & Associates. (1980). *Toward reform of program evaluation*. San Francisco, CA: Jossey-Bass.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measures*. Hoboken, NJ: Wiley.

Cronbach, L. J., & Snow, R. E. (1969). *Individual differences in learning ability as a function of instructional variables*. Redwood City, CA: Stanford University Press.

Cronbach, L. J., & Snow, R. E. (1981). *Aptitudes and instructional methods: A handbook for research on interactions*. New York, NY: Irvington.

Cryer, J. D., & Chan, K.-S. (2010). *Time series analysis: With applications in R* (2nd ed.). Hoboken, NJ: Wiley.

Cullen, A. E. (2009). *The politics and consequences of stakeholder participation in international development evaluations*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Cullen, A. E., & Coryn, C.L.S. (2011). Forms and functions of participatory evaluation: A review of the empirical and theoretical literature. *Journal of MultiDisciplinary Evaluation*, *7*(16), 32–47.

Cullen, A. E., Coryn, C.L.S., & Rugh, J. (2011). The politics and consequences of including stakeholders in international development evaluations. *American Journal of Evaluation*, *32*, 345–361.

Daigneault, P.-M., & Jacob, S. (2009). Toward accurate measurement of participation: Rethinking the conceptualization and operationalization of participatory evaluation. *American Journal of Evaluation*, *30*, 330–348.

Datta, L.-E. (1999). CIRCE's demonstration of a close-to-ideal evaluation in a less-than-ideal world. *American Journal of Evaluation*, *20*, 345–354.

Datta, L.-E. (2003). Important questions, intriguing method, incomplete answers. In C. A. Christie (Ed.), *The practice-theory relationship in evaluation* (pp. 37–46). New Directions for Evaluation, no. 97. San Francisco, CA: Jossey-Bass.

Davey, J. W., Gugiu, P. C., & Coryn, C.L.S. (2010). Quantitative methods for estimating the reliability of qualitative data. *Journal of MultiDisciplinary Evaluation*, *6*(13), 140–162.

Davidson, E. J. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation*. Thousand Oaks, CA: Sage.

Davidson E. J. (2007). Unlearning some of our social scientist habits. *Journal of MultiDisciplinary Evaluation*, *4*(8), iii–vi.

Davidson, E. J. (2011, November). *The rubric revolution: Evaluative blending of mixed method evidence*. Paper presented at the annual meeting of the American Evaluation Association, Anaheim, CA.

Davis, H. R., & Salasin, S. E. (1975). The utilization of evaluation. In E. L. Struening & M. Guttentag (Eds.), *Handbook of evaluation research* (pp. 621–665). Thousand Oaks, CA: Sage.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Dellarocas, C. (2003). The digitization of word of mouth: Promises and challenges of online feedback mechanisms. *Management Science*, *49*, 1407–1424.

Denny, T. (1978). *Storytelling and educational understanding* (Occasional Paper Series, Paper #12). Kalamazoo: Western Michigan University, Evaluation Center.

Denny, T. (2011). Storytelling and educational understanding. *Journal of MultiDisciplinary Evaluation*, *7*(15), 258–271.

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2005). *The Sage handbook of qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.

Dewey, J. (1934). *Art as experience*. New York, NY: Milton Balch.

Dewey, J. D., Montrosse, B. E., Schröter, D. C., Sullins, C. D., & Mattox, J. R., II. (2008). Evaluator competencies: What's taught versus what's sought. *American Journal of Evaluation*, *29*, 268–287.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* (3rd ed.). Hoboken, NJ: Wiley.

Donaldson, S. I. (2007). *Program theory–driven evaluation science*. Mahwah, NJ: Lawrence Erlbaum Associates.

Donaldson, S. I. (2013). *The future of evaluation in society: A tribute to Michael Scriven*. Charlotte, NC: Information Age.

Donaldson, S. I., & Christie, C. A. (2005). The 2005 Claremont debate: Lipsey versus Scriven. Determining causality in program evaluation and applied research: Should experimental evidence be the gold standard? *Journal of MultiDisciplinary Evaluation*, *2*(3), 60–77.

Donaldson, S. I., Christie, C. A., & Mark, M. M. (Eds.). (2009). *What counts as credible evidence in applied research and program evaluation practice?* Thousand Oaks, CA: Sage.

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

Donaldson, S. I., Patton, M. Q., Fetterman, D. M., & Scriven, M. (2010). The 2009 Claremont debates: The promise and pitfalls of utilization-focused and empowerment evaluation. *Journal of MultiDisciplinary Evaluation*, *6*(13), 15–57.

Eisner, E. W. (1975). *The perceptive eye: Toward the reformation of educational evaluation.* Stanford, CA: Stanford Evaluation Consortium.

Eisner, E. W. (1983). Educational connoisseurship and criticism: Their form and functions in educational evaluation. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 335–348). Norwell, MA: Kluwer.

Eisner, E. W. (1985). *The art of educational evaluation: A personal view*. London, UK: Falmer Press.

Eisner, E. (1991). 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 11* (pp. 169–187). Chicago, IL: University of Chicago Press.

Eisner, E. W. (1998). *The enlightened eye: Qualitative inquiry and the enhancement of educational practice*. Upper Saddle River, NJ: Merrill.

Eisner, E. W. (2004). The roots of connoisseurship and criticism: A personal journey. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 196–202). Thousand Oaks, CA: Sage.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge, UK: Cambridge University Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Evaluation Research Society Standards Committee. (1982). Evaluation Research Society standards for program evaluation. In P. H. Rossi (Ed.), *Standards for evaluation practice* (pp. 7–19). New Directions for Program Evaluation, no. 15. San Francisco, CA: Jossey-Bass.

Evergreen, S.D.H. (2010, November). *Why evaluators need graphic design skills*. Paper presented at the annual meeting of the American Evaluation Association, San Antonio, TX.

Evergreen, S.D.H. (2011). *Death by boredom: The role of visual processing theory in written evaluation communication*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Evergreen, S.D.H., & Cullen, A. (2010). Moving to genuine: Credible cultural competence. *Journal of MultiDisciplinary Evaluation*, *6*(13), 130–139.

Evergreen, S.D.H., & Robertson, K. (2010). How do evaluators communicate cultural competence? Indications of cultural competence through an examination of the American Evaluation Association's Career Center. *Journal of MultiDisciplinary Evaluation*, *6*(14), 58–67.

Evers, J. (1980). *A field study of goal-based and goal-free evaluation techniques*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Fetterman, D. M. (1984). *Ethnography in educational evaluation*. Thousand Oaks, CA: Sage.

Fetterman, D. M. (1994). Empowerment evaluation. *Evaluation Practice*, *15*, 1–15.

Fetterman, D. M. (1998). *Ethnography: Step by step* (2nd ed.). Thousand Oaks, CA: Sage.

Fetterman, D. M. (2001). *Foundations of empowerment evaluation: Step by step*. Thousand Oaks, CA: Sage.

Fetterman, D. M. (2004). Branching out or standing on a limb: Looking to our roots for insight. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 319–330). Thousand Oaks, CA: Sage.

Fetterman, D. M. (2005). Empowerment evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 125–129). Thousand Oaks, CA: Sage.

Fetterman, D. M., & Wandersman, A. (Eds.). (2005). *Empowerment evaluation principles in practice*. New York, NY: Guilford Press.

Few, S. (2009). *Now you see it: Simple visualization techniques for quantitative analysis*. Oakland, CA: Analytics Press.

Fielding, N. G., & Lee, R. M. (1991). *Using computers in qualitative research*. Thousand Oaks, CA: Sage.

Fielding, N. G., & Lee, R. M. (1998). *Computer analysis and qualitative research*. Thousand Oaks, CA: Sage.

Finkelstein, A., Taubman, S. Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., . . . Oregon Health Study Group. (2011). *The Oregon health insurance experiment: Evidence from the first year* (Working Paper 17190). Cambridge, MA: National Bureau of Economic Research.

Finn, C. E., Stevens, F. I., Stufflebeam, D. L., & Walberg, H. J. (1997). A meta-evaluation. *International Journal of Educational Research*, *27*, 159–174.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, UK: Oliver & Boyd.

Fisher, R. A. (1951). *The design of experiments* (6th ed.). New York, NY: Hafner.

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2011). *Program evaluation: Alternative approaches and practical guidelines* (4th ed.). Upper Saddle River, NJ: Pearson.

Flay, B. R., Biglan, A., Boruch, R. F., González Castro, F., Gottfredson, D., Kellam, S., . . . Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science*, *6*, 151–175.

Fleck, A. C. (1961). Evaluation as a logical process. *Canadian Journal of Public Health*, *52*, 185–191.

Flexner, A. (1910). *Medical education in the United States and Canada*. Bethesda, MD: Science and Health.

Flinders, D. J., & Eisner, E. W. (2000). Educational criticism as a form of qualitative inquiry. In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed., pp. 195–208). Norwell, MA: Kluwer.

Fowler, F. J., Jr. (1995). *Improving survey questions: Design and evaluation*. Applied Social Research Methods Series, Vol. 38. Thousand Oaks, CA: Sage.

Fournier, D. M. (1995). Establishing evaluative conclusions: A distinction between general and working logic. In D. M. Fournier (Ed.), *Reasoning in evaluation: Inferential links and leaps* (pp. 15–32). New Directions for Evaluation, no. 68. San Francisco, CA: Jossey-Bass.

Freeman, M. (Ed.). (2010). *Critical social theory and evaluation practice*. New Directions for Evaluation, no. 127. San Francisco, CA: Jossey-Bass.

Funnel, S. C., & Rogers, P. J. (2011). *Purposeful program theory: Making effective use of theories of change and program logic models*. San Francisco, CA: Jossey-Bass.

Gally, J. (1984, April). *The evaluation component*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Garaway, G. B. (1995). Participatory evaluation. *Studies in Educational Evaluation*, *21*, 85–102.

Gilligan, C. (1982). *In a different voice*. Cambridge, MA: Harvard University Press.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory*. Chicago, IL: Aldine.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, *18*, 519–521.

Glass, G. V. (1975). A paradox about excellence of schools and the people in them. *Educational Researcher*, *4*(3), 9–13.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–10.

Glass, G. V., Willson, V. L., & Gottman, J. M. (2008). *Design and analysis of time-series experiments*. Charlotte, NC: Information Age.

Goldstein, H. (1987). *Multilevel models in educational and social research*. New York, NY: Oxford University Press.

Grant, S., & Humphries, S. (2006). Critical evaluation of appreciative inquiry: Bridging an apparent paradox. *Action Research*, *4*, 401–418.

Grasso, P. G. (1999). Meta-evaluation of an evaluation of Reader Focused Writing for the Veterans Benefits Administration. *American Journal of Evaluation*, *20*, 355–371.

Green, L. W., & Kreuter, M. W. (1991). *Health promotion planning: An educational and environmental approach* (2nd ed.). Mountain View, CA: Mayfield.

Greene, J. C. (1988a). Communication of results and utilization in participatory program evaluation. *Evaluation and Program Planning*, *11*, 341–351.

Greene, J. C. (1988b). Stakeholder participation and utilization in program evaluation. *Evaluation Review*, *12*, 91–116.

Greene, J. C. (2004). The educative evaluator: An interpretation of Lee J. Cronbach's vision of evaluation. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 169–180). Thousand Oaks, CA: Sage.

Greene, J. C. (2007). *Mixed methods in social inquiry*. San Francisco, CA: Jossey-Bass.

Greene, J. C. (2008). Is mixed methods social inquiry a distinctive methodology? *Journal of Mixed Methods Research*, *2*, 7–22.

Greene, J. C., & Abma, T. A. (Eds.). (2001). *Responsive evaluation*. New Directions for Evaluation, no. 92. San Francisco, CA: Jossey-Bass.

Guba, E. G. (1966, October). *A study of Title III activities: Report on evaluation*. Paper presented at the National Institute for the Study of Educational Change, Indiana University, Bloomington.

Guba, E. G. (1969). The failure of educational evaluation. *Educational Technology*, *9*(5), 29–38.

Guba, E. G. (1977). *Educational evaluation: The state of the art*. Keynote address at the annual meeting of the Evaluation Network, St. Louis, MO.

Guba, E. G. (1978). *Toward a methodology of naturalistic inquiry in educational evaluation*. Los Angeles: University of California, Center for the Study of Evaluation.

Guba, E. G. (1990). *The paradigm dialog*. Thousand Oaks, CA: Sage.

Guba, E. G., & Lincoln, Y. S. (1981). *Effective evaluation*. San Francisco, CA: Jossey-Bass.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Thousand Oaks, CA: Sage.

Guba, E. G., & Stufflebeam, D. L. (1968). Evaluation: The process of stimulating, aiding, and abetting insightful action. In R. Ingle & W. Gephart (Eds.), *Problems in the training of educational researchers* (pp. 1–35). Bloomington, IN: Phi Delta Kappa.

Guba, E. G., & Stufflebeam, D. L. (1970, June). *Strategies for the institutionalization of the CIPP evaluation model*. Keynote address given at the 11th Phi Delta Kappa Symposium on Educational Research, Columbus, OH.

Gubrium, J. F., & Holstein, J. A. (Eds.). (2001). *Handbook of interview research: Context and method*. Thousand Oaks, CA: Sage.

Gugiu, P. C. (2007). The logic of summative confidence. *Journal of MultiDisciplinary Evaluation*, *4*(8), 1–15.

Gugiu, P. C. (2011). *Summative confidence*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Guilford, J. P. (1936). *Psychometric methods*. New York, NY: McGraw-Hill.

Gullickson, A. R., & Stufflebeam, D. L. (2001). *Feedback workshop checklist*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/feedbackworkshop.pdf

Gunter, H., & Rayner, S. (2007). Modernizing the school workforce in England: Challenging transformation and leadership? *Leadership*, *3*, 47–64.

Habashi, J., & Worley, J. (2009). Child geopolitical agency: A mixed methods case study. *Journal of Mixed Methods Research*, *3*, 42–64.

Hamilton, D. (2005). Illuminative evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 191–194). Thousand Oaks, CA: Sage.

Hammond, R. L. (1967). *Evaluation at the local level*. Address to the Miller Committee for the National Study of Elementary and Secondary Education Act Title III, Washington, DC.

Hammond, R. L. (1972). *Evaluation at the local level*. Tucson, AZ: EPIC Evaluation Center.

Hastings, T. (1976). *A portrayal of the changing evaluation scene*. Keynote address at the annual meeting of the Evaluation Network, St. Louis, MO.

Heberger, A. E., Christie, C. A., & Alkin, M. C. (2010). A bibliometric analysis of the academic influences of and on evaluation theorists' published works. *American Journal of Evaluation*, *31*, 24–44.

Hendricks, M., & Handley, E. A. (1990). Improving the recommendations from evaluation studies. *Evaluation and Program Planning*, *13*, 109–117.

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, *18*(1), 38–52.

Henry, G. T. (1990). *Practical sampling*. Applied Social Research Methods Series, Vol. 31. Thousand Oaks, CA: Sage.

Henry, G. T. (2005). Realist evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 359–362). Thousand Oaks, CA: Sage.

Henry, G. T., Julnes, G., & Mark, M. M. (Eds.). (1998). *Realist evaluation: An emerging theory in support of practice*. New Directions for Evaluation, no. 78. San Francisco, CA: Jossey-Bass.

Herzog, E. (1959). *Some guidelines for evaluative research*. Washington, DC: U.S. Department of Health, Education, and Welfare.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics* (5th ed.). Boston, MA: Houghton Mifflin.

Hobson, K. A., Mateu, P., Coryn, C.L.S., & Graves, C. D. (2012). Measles, mumps, and rubella vaccines and diagnoses of autism spectrum disorders among children: A meta-analysis. *World Medical & Health Policy*, *4*(3), 1–14.

Hodgkin, S. (2008). Telling it all: A story of women's social capital using a mixed methods approach. *Journal of Mixed Methods Research*, *2*, 296–316.

Hofstetter, C., & Alkin, M. C. (2003). Evaluation use revisited. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 197–222). Norwell, MA: Kluwer.

Hopkins, K. D., & Glass, G. V. (1978). *Basic statistics for the behavioral sciences*. Upper Saddle River, NJ: Prentice Hall.

Horn, J. (2001). *A checklist for developing and evaluating evaluation budgets*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/evaluationbudgets.pdf

House, E. R. (Ed.). (1973). *School evaluation: The politics and process*. Berkeley, CA: McCutchan.

House, E. R. (1980). *Evaluating with validity*. Thousand Oaks, CA: Sage.

House, E. R. (1983). Assumptions underlying evaluation models. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 45–64). Norwell, MA: Kluwer.

House, E. R. (1993). *Professional evaluation: Social impact and political consequences*. Thousand Oaks, CA: Sage.

House, E. R. (2004). Intellectual history in evaluation. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 218–224). Thousand Oaks, CA: Sage.

House, E. R. (2005). Deliberative democratic evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 104–108). Thousand Oaks, CA: Sage.

House, E. R., & Howe, K. R. (1998). *Deliberative democratic evaluation in practice*. Boulder: University of Colorado.

House, E. R., & Howe, K. R. (2000a). Deliberative democratic evaluation. In K. E. Ryan & L. DeStefano (Eds.), *Evaluation as a democratic process: Promoting inclusion, dialogue, and deliberation* (pp. 3–12). New Directions for Evaluation, no. 85. San Francisco, CA: Jossey-Bass.

House, E. R., & Howe, K. R. (2000b). *Deliberative democratic evaluation checklist*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/dd_checklist.PDF

House, E. R., & Howe, K. R. (2000c). Deliberative democratic evaluation in practice. In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed., pp. 409–421). Norwell, MA: Kluwer.

House, E. R., & Howe, K. R. (2003). Deliberative democratic evaluation. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 79–100). Norwell, MA: Kluwer.

Hughes, M., & Kushner, S. (2005). Accreditation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 4–7). Thousand Oaks, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.

Jaeger, R. M. (1990). *Statistics: A spectator sport* (2nd ed.). Thousand Oaks, CA: Sage.

James, G. (1958). Research by local health departments—problems, methods, results. *American Journal of Public Health*, *48*, 354–379.

Jang, S. (2000). The appropriateness of Joint Committee standards in non-Western settings: A case study of South Korea. In C. Russon (Ed.), *The program evaluation standards in international settings* (pp. 41–59). Kalamazoo: Western Michigan University, Evaluation Center.

Janz, N. K., & Becker, M. H. (1984). The health belief model: A decade later. *Health Education Quarterly*, *11*, 1–47.

Johnson, K., Greenseid, L. O., Toal, S. A., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation*, *30*, 377–410.

Joint Committee on Standards for Educational Evaluation. (1981). *Standards for evaluations of educational programs, projects, and materials*. New York, NY: McGraw-Hill.

Joint Committee on Standards for Educational Evaluation. (1988). *The personnel evaluation standards*. Thousand Oaks, CA: Corwin Press.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Joint Committee on Standards for Educational Evaluation. (2003). *The student evaluation standards*. Thousand Oaks, CA: Corwin Press.

Joint Committee on Standards for Educational Evaluation. (2009). *The personnel evaluation standards: How to assess systems for evaluating educators* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Joint Committee on Standards for Educational Evaluation. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Jones, L. V. (2003). National assessment in the United States: The evolution of a nation's report card. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 883–904). Norwell, MA: Kluwer.

Jones, S. C., & Worthen, B. R. (1999). AEA members' opinions concerning evaluator certification. *American Journal of Evaluation*, *20*, 495–506.

Jorgensen, D. L. (1989). *Participant observation: A methodology for human studies*. Thousand Oaks, CA: Sage.

Kaplan, A. (1964). *The conduct of inquiry*. San Francisco, CA: 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* (pp. 21–38). New Directions for Evaluation, no. 77. San Francisco, CA: Jossey-Bass.

Kee, J. E. (1995). Benefit-cost analysis in program evaluation. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (pp. 456–488). San Francisco, CA: Jossey-Bass.

Kellaghan, T. (1982). Los sistemas escalates coma objecto de evaluacion. In D. L. Stufflebeam, T. Kellaghan, & B. Alvarez (Eds.), *La evaluacion educativa*. Bogota, Columbia: Pontificia Universidad Javeriana.

Kellaghan, T., Madaus, G., & Airasian, P. (1982). *The effects of standardized testing*. Norwell, MA: Kluwer.

Kellaghan, T., & Stufflebeam, D. L. (Eds.). (2003). *International handbook of educational evaluation*. Norwell, MA: Kluwer.

Kemple, J. J., with Scott-Clayton, J. (2004). *Career academies: Impacts on labor market outcomes and educational attainment*. New York, NY: MDRC.

Kerlinger, F. N. (1986). *Foundations of behavioral research* (3rd ed.). New York, NY: Holt, Rinehart and Winston.

Kidder, L., & Fine, M. (1987). Qualitative and quantitative methods: When stories converge. In M. M. Mark & R. Shotland (Eds.), *Multiple methods in program evaluation* (pp. 57–75). New Directions for Program Evaluation, no. 35. San Francisco, CA: Jossey-Bass.

King, J. A. (1998). Making sense of participatory evaluation practice. In E. Whitmore (Ed.), *Understanding and practicing participatory evaluation* (pp. 57–67). New Directions for Evaluation, no. 80. San Francisco, CA: Jossey-Bass.

King, J. A. (2004). Tikkun Olam: The roots of participatory evaluation. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 331–342). Thousand Oaks, CA: Sage.

King, J. A. (2005). Participatory evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 291–294). Thousand Oaks, CA: Sage.

King, J. A. (2007). Making sense of participatory evaluation practice. In S. Mathison (Ed.), *Enduring issues in evaluation: The 20th anniversary of the collaboration between NDE and AEA* (pp. 83–105). New Directions for Evaluation, no. 114. San Francisco, CA: 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.

Kirst, M. W. (1990). *Accountability: Implications for state and local policymakers*. Washington, DC: U.S. Department of Education, Information Services, Office of Educational Research and Improvement.

Kish, L. (1965). *Survey sampling*. Hoboken, NJ: Wiley.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Kline, R. B. (2008). *Becoming a behavioral science researcher: A guide to producing research that matters*. New York, NY: Guilford Press.

Klineberg, O. (1955). The problem of evaluation. *International Social Science Bulletin*, *7*, 347–362.

Koleci, X., Coryn, C.L.S., Hobson, K. A., & Keci, R. (2011). Probability sampling designs for veterinary epidemiology. *Albanian Journal of Agricultural Sciences*, *3*(10), 1–16.

Koretz, D. (1996a). Using student assessments for educational accountability. In R. Hanushek (Ed.), *Improving the performance of America's schools* (pp. 171–196). Washington, DC: National Academies Press.

Koretz, D. (1996b). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS)*. Santa Monica, CA: RAND.

Kruse, R. L., Alper, B. S., Reust, C., Stevermer, J. J., Shannon, S., & Williams, R. H. (2002). Intention-to-treat analysis: Who is in? Who is out? *The Journal of Family Practice*, *51*, 969–971.

Kvale, S. (1996). *InterViews: An introduction to qualitative research interviewing*. Thousand Oaks, CA: Sage.

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*, 9–23.

LeCompte, M. D., & Goetz, J. P. (1982). Problems of reliability and validity in ethnographic research. *Review of Educational Research*, *52*, 31–60.

Leninger, M. (Ed.). (1985). *Qualitative research methods in nursing*. Orlando, FL: Grune & Stratton.

Lessinger, L. M. (1970). *Every kid a winner: Accountability in education*. New York, NY: Simon & Schuster.

Levin, B. (1993). Collaborative research in and with organizations. *Qualitative Studies in Education*, *6*, 331–340.

Levin, H. M. (1983). *Cost-effectiveness: A primer*. Thousand Oaks, CA: Sage.

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications* (2nd ed.). Thousand Oaks, CA: Sage.

Levine, M. (1974). Scientific method and the adversary model: Some preliminary thoughts. *American Psychologist*, *29*, 666–677.

Lewin, K. (1952). *Field theory in social science: Selected theoretical papers*. London, UK: Tavistock.

Lincoln, Y. S. (2003). Constructivist knowing, participatory ethics and responsive evaluation: A model for the 21st century. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 69–78). Norwell, MA: Kluwer.

Lincoln, Y. S. (2005). Fourth-generation evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 161–164). Thousand Oaks, CA: Sage.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Thousand Oaks, CA: Sage.

Lincoln, Y. S., & Guba, E. G. (2004). The roots of fourth generation evaluation. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 225–242). Thousand Oaks, CA: Sage.

Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education*. Boston, MA: Houghton Mifflin.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15–21.

Lipsey, M. W., & Hurley, S. M. (2009). Design sensitivity: Statistical power for applied experimental research. In L. Bickman & D. J. Rog (Eds.), *The Sage handbook of applied social research methods* (2nd ed., pp. 44–76). Thousand Oaks, CA: Sage.

Lipsey, M. W., Rossi, P. H., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Applied Social Research Methods Series, Vol. 49. Thousand Oaks, CA: Sage.

Lohr, S. L. (2010). *Sampling: Design and analysis* (2nd ed.). Belmont, CA: Thompson.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mabry, L. (2003). In living color: Qualitative methods in educational evaluation. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 167–188). Norwell, MA: Kluwer.

MacDonald, B. (1975). Evaluation and the control of education. In D. Tawney (Ed.), *Evaluation: The state of the art* (pp. 125–136). London, UK: Schools Council.

MacNeil, C. (2005). Critical theory evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 92–94). Thousand Oaks, CA: Sage.

Madaus, G. F. (2004). Ralph W. Tyler's contribution to program evaluation. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 69–79). Thousand Oaks, CA: Sage.

Madaus, G. F., Scriven, M., & Stufflebeam, D. L. (Eds.). (1983). *Evaluation models: Viewpoints on educational and social services evaluation*. Norwell, MA: Kluwer.

Madaus, G. F., & Stufflebeam, D. L. (1988). *Educational evaluation: The classical writings of Ralph W. Tyler*. Norwell, MA: Kluwer.

Mafukidze-Trent, T. (2009). *Metaevaluation of HIV/AIDS prevention intervention evaluations in Sub-Saharan Africa with a specific emphasis on implications for women and girls*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Margolis, E., & Pauwels, L. (Eds.). (2011). *The Sage handbook of visual research methods*. Thousand Oaks, CA: Sage.

Mark, M. M., Donaldson, S. I., & Campbell, B. (2011). *Social psychology and evaluation*. New York, NY: Guilford Press.

Mark, M. M., Henry, G. T., & Julnes, G. (2000). *Evaluation: An integrative framework for understanding, guiding, and improving policies and programs*. San Francisco, CA: Jossey-Bass.

Mark, M. M., & Shotland, R. L. (1985). Stakeholder-based evaluation and value judgments. *Evaluation Review*, *9*, 605–626.

Mathison, S. (Ed.). (2005a). *Encyclopedia of evaluation*. Thousand Oaks, CA: Sage.

Mathison, S. (2005b). Mertens, Donna M. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 247–248). Thousand Oaks, CA: Sage.

McDavid, J. C., & Hawthorn, L.R.L. (2006). *Program evaluation and performance measurement: An introduction to practice*. Thousand Oaks, CA: Sage.

Mehrens, W. A. (1972). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice*, *11*(1), 3–10.

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. (2003). The inclusive view of evaluation: Visions for the new millennium. In S. I. Donaldson & M. Scriven (Eds.), *Evaluating social programs and problems: Visions for the new millennium* (pp. 87–104). Mahwah, NJ: Lawrence Erlbaum Associates.

Mertens, D. M. (2005a). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods* (2nd ed.). Thousand Oaks, CA: Sage.

Mertens, D. M. (2005b). Transformative paradigm. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 422–423). Thousand Oaks, CA: Sage.

Mertens, D. M. (2007a). Transformative considerations: Inclusion and social justice. *American Journal of Evaluation*, *28*, 86–90.

Mertens, D. M. (2007b). Transformative paradigm: Mixed methods and social justice. *Journal of Mixed Methods Research*, *1*, 212–225.

Mertens, D. M. (2009). *Transformative research and evaluation*. New York, NY: Guilford Press.

Mertens, D. M., Farley, J., Singleton, P., & Madison, A. (1994). Diverse voices in evaluation practice: Feminists, minorities, and persons with disabilities. *Evaluation Practice*, *15*, 123–129.

Mertens, D. M., Harris, R., Holmes, H., & Brandt, S. (2007). *Project SUCCESS: Summative evaluation report*. Washington, DC: Gallaudet University.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(3), 13–23.

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

Miles, M. B., & Huberman, A. M. (1984). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.

Miller, H. L. (Ed.). (1997). The New York City Public Schools integrated learning systems project: Evaluation and meta-evaluation [Special issue]. *International Journal of Educational Research*, *27*(2).

Miller, R. L. (2010). Developing standards for empirical examinations of evaluation theory. *American Journal of Evaluation*, *31*, 390–399.

Miller, R. L., & Campbell, R. (2006). Taking stock of empowerment evaluation: An empirical review. *American Journal of Evaluation*, *27*, 296–319.

Millett, R. (1995). *W. K. Kellogg Foundation cluster evaluation model of evolving practices*. Battle Creek, MI: W. K. Kellogg Foundation.

Millett, R. (1996). Empowerment evaluation and the W. K. Kellogg Foundation. In D. M. Fetterman, A. J. Kaftarian, & A. Wandersman (Eds.), *Empowerment evaluation: Knowledge and tools for self-assessment and accountability* (pp. 65–76). Thousand Oaks, CA: Sage.

Morell, J. A. (2010). *Evaluation in the face of uncertainty: Anticipating surprise and responding to the inevitable*. New York, NY: Guilford Press.

Morris, M. (2003). Ethical considerations in evaluation. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 303–328). Norwell, MA: Kluwer.

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

Morris, M. (2011). The good, the bad, and the evaluator: 25 years of AJE ethics. *American Journal of Evaluation*, *32*, 134–151.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children*, *5*(2), 113–127.

Mullen, P. D., Hersey, J., & Iverson, D. C. (1987). Health behavior models compared. *Social Science and Medicine*, *24*, 973–981.

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. New York, NY: Oxford University Press.

Nader, R. (1965). *Unsafe at any speed: The designed-in dangers of the American automobile*. New York, NY: Grossman.

National Science Foundation. (1997). *User-friendly handbook for mixed method evaluations*. Arlington, VA: Author.

Nave, B., Miech, E. J., & Mosteller, F. (2000). A rare design: The role of field trials in evaluating school practices. In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed., pp. 145–162). Norwell, MA: Kluwer.

Nevo, D. (1974). *Evaluation priorities of students, teachers, and principals*. Unpublished doctoral dissertation, Ohio State University, Columbus.

Nevo, D. (1993). The evaluation minded school: An application of perceptions from program evaluation. *Evaluation Practice*, *14*, 39–47.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, *231*, 289–337.

Nowakowski, J. A. (1974). *Handbook for traveling observers*. Kalamazoo: Western Michigan University, Evaluation Center.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Orris, M. J. (1989). *Industrial applicability of the Joint Committee's personnel evaluation standards*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Osgood, C. E., Suci, J. G., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

O'Sullivan, R. G., & D'Agostino, A. (2002). Promoting evaluation through collaboration: Findings from community-based programs for young children and their families. *Evaluation: The International Journal of Theory, Research and Practice*, *8*, 372–387.

Owen, J. M. (2004). Evaluation forms: Toward an inclusive framework for evaluation practice. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 356–369). Thousand Oaks, CA: Sage.

Owen, J. M. (2006). *Program evaluation: Forms and approaches* (3rd ed.). New York, NY: Guilford Press.

Owen, J. M., & Rogers, P. J. (1999). *Program evaluation: Forms and approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Owens, T. (1973). Educational evaluation by adversary proceeding. In E. House (Ed.), *School evaluation: The politics and process* (pp. 295–305). Berkeley, CA: McCutchan.

Parlett, M., & Hamilton, D. (1972). *Evaluation as illumination: A new approach to the study of innovatory programs*. Edinburgh, UK: University of Edinburgh, Centre for Research in the Educational Sciences.

Patton, M. Q. (1980). *Qualitative evaluation methods*. Thousand Oaks, CA: Sage.

Patton, M. Q. (1982). *Practical evaluation*. Thousand Oaks, CA: Sage.

Patton, M. Q. (1984). An alternative evaluation approach for the problem-solving training program: A utilization-focused evaluation process. *Evaluation and Program Planning*, *7*, 189–192.

Patton, M. Q. (1987). How to use qualitative methods in evaluation. In J. L. Herman (Ed.), *Program evaluation kit* (2nd ed.; Vol. 4). Thousand Oaks, CA: Sage.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (1994). Developmental evaluation. *Evaluation Practice*, *15*, 311–319.

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

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2000). Utilization-focused evaluation. In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed., pp. 425–438). Norwell, MA: Kluwer.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2003). Utilization-focused evaluation. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 223–244). Norwell, MA: Kluwer.

Patton, M. Q. (2004). The roots of utilization-focused evaluation. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 276–292). 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*, 408–414.

Patton, M. Q. (2005b). Utilization-focused evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 429–432). Thousand Oaks, CA: Sage.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2010). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. New York, NY: Guilford Press.

Patton, M. Q. (2012). *Essentials of utilization-focused evaluation*. Thousand Oaks, CA: Sage.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. London, UK: Sage.

Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal*, *2*(2228), 1243–1246.

Persaud, N. (2007). *Conceptual and practical analysis of costs and benefits: Developing a cost-analysis tool for practical program evaluation*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Peters, T. J., & Waterman, R. H. (1982). *In search of excellence: Lessons from America's best-run companies*. New York, NY: Warner Books.

Platt, J. (1992). Case study in American methodological thought. *Current Sociology*, *40*(1), 17–48.

Popham, W. J. (1969). Objectives and instruction. In R. Stake (Ed.), *Instructional objectives* (pp. 65–90). Skokie, IL: Rand McNally.

Popham, W. J. (1971). *Criterion-referenced measurement*. Upper Saddle River, NJ: Educational Technology.

Popham, W. J., & Carlson, D. (1977). Deep dark deficits of the adversary evaluation model. *Educational Researcher*, *6*(6), 3–6.

Posavac, E. J., & Carey, R. G. (2003). *Program evaluation: Methods and case studies* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Preskill, H. (1994). Evaluation's role in enhancing organizational learning. *Evaluation and Program Planning*, *17*, 291–297.

Preskill, H. (2005). Appreciative inquiry. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 18–19). Thousand Oaks, CA: Sage.

Preskill, H., & Catsambas, T. T. (2006). *Reframing evaluation through appreciative inquiry*. Thousand Oaks, CA: Sage.

Preskill, H., & Coghlan, A. T. (Eds.). (2003). *Using appreciative inquiry in evaluation*. New Directions for Evaluation, no. 100. San Francisco, CA: Jossey-Bass.

Preskill, H., & Russ-Eft, D. (2005). *Building evaluation capacity: 72 activities for teaching and training*. Thousand Oaks, CA: Sage.

Preskill, H., & Torres, R. T. (1999a). Building capacity for organizational learning through evaluative inquiry. *Evaluation*, *5*, 42–60.

Preskill, H., & Torres, R. T. (1999b). *Evaluative inquiry for learning in organizations*. Thousand Oaks, CA: Sage.

Prochaska, J. O., & DiClemente, C. C. (1992). Stages of change in the modification of problem behaviors. In M. Hersen, R. M. Eisler, & P. M. Miller (Eds.), *Progress in behavior modification* (Vol. 28, pp. 183–218). Sycamore, IL: Sycamore.

Provus, M. N. (1969). *Discrepancy evaluation model*. Pittsburgh, PA: Pittsburgh Public Schools.

Provus, M. N. (1971). *Discrepancy evaluation*. Berkeley, CA: McCutchan.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Advanced Quantitative Techniques in the Social Sciences Series, Vol. 1. Thousand Oaks, CA: Sage.

Raudenbush, S. W., Martinez, A., & Spybrook, J. K. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, *29*(1), 5–29.

Reichardt, C. S. (2005). Quasi-experimental design. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 351–355). Thousand Oaks, CA: Sage.

Reed, M. (1989). *WMU traveling observer handbook* (5th ed.). Kalamazoo: Western Michigan University, Evaluation Center.

Reinhard, D. (1972). *Methodology development for input evaluation using advocate and design teams*. Unpublished doctoral dissertation, Ohio State University, Columbus.

Resnick, P., Zeckhauser, R., Friedman, E., & Kuwabara, K. (2000). Reputation systems. *Communications of the ACM*, *43*(12), 45–48.

Rippey, R. M. (Ed.). (1973). *Studies in transactional evaluation*. Berkeley, CA: McCutchan.

Risley, J. S. (2007). *Legislative program evaluation conducted by state legislatures in the United States*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Robinson, T. T., & Cousins, J. B. (2004). Internal participatory evaluation as an organizational learning system: A longitudinal case study. *Studies in Educational Evaluation*, *30*, 1–22.

Rodriguez-Campos, L. (2005). *Collaborative evaluations: A step-by-step model for the evaluator*. Tarmac, FL: Lumina Press.

Rogers, P. J. (2000). Program theory: Not whether programs work but how they work. In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed., pp. 209–232). Norwell, MA: Kluwer.

Rogers, P. J. (2008). Using programme theory to evaluate complicated and complex aspects of interventions. *Evaluation*, *14*, 29–48.

Rossi, P. H., & Freeman, H. E. (1993). *Evaluation: A systematic approach* (5th ed.). Thousand Oaks, CA: Sage.

Rossi, P. H., Freeman, H. E., & Rosenbaum, S. (1979). *Evaluation: A systematic approach*. Thousand Oaks, CA: Sage.

Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (1999). *Evaluation: A systematic approach* (6th ed.). Thousand Oaks, CA: Sage.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322–331.

Russon, C. (2005). Cluster evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 66–67). Thousand Oaks, CA: Sage.

Ryan, K. (2004). Serving public interests in educational accountability: Alternative approaches to democratic evaluation. *American Journal of Evaluation*, *25*, 443–461.

Ryan, K. (2005). Democratic evaluation approaches for equity and inclusion. *The Evaluation Exchange*, *11*(3), 2–3.

Ryan, K., Greene, J., Lincoln, Y., Mathison, S., & Mertens, D. M. (1998). Advantages and challenges of using inclusive evaluation approaches in evaluation practice. *American Journal of Evaluation*, *19*, 101–122.

Sandberg, J. (1986). *Alabama educator inservice traveling observer handbook*. Kalamazoo: Western Michigan University, Evaluation Center.

Sanders, J. R. (1992). *Evaluating school programs*. Thousand Oaks, CA: Sage.

Sanders, J. R. (1995). Standards and principles. In W. R. Shadish, D. L. Newman, M. A. Scheirer, & C. Wye (Eds.), *Guiding principles for evaluators* (pp. 47–53). New Directions for Program Evaluation, no. 66. San Francisco, CA: Jossey-Bass.

Sanders, J. R. (1997). Cluster evaluation. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook* (pp. 396–404). Thousand Oaks, CA: Sage.

Sanders, W. L. (1989). *Using customized standardized tests*. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Sanders, W. L., & Horn, S. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, *8*, 299–311.

Sasaki, R. (2008). *Metaevaluation by formal evaluation theory of aid evaluation work*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Scheaffer, R. L., Mendenhall, W., III, Ott, R. L., & Gerow, K. G. (2012). *Elementary survey sampling* (7th ed.). Belmont, CA: Thompson.

Schröter, D. C., Coryn, C.L.S., & Youker, B. W. (2006). *Evaluation of the 2005 Kalamazoo Public Schools Middle School Summer Enrichment Program: Synthesis*. Kalamazoo: Western Michigan University, Evaluation Center.

Schwandt, T. A. (1984). *An examination of alternative models for socio-behavioral inquiry*. Unpublished doctoral dissertation, Indiana University, Bloomington.

Schwandt, T. A. (1989). Recapturing moral discourse in evaluation. *Educational Researcher*, *18*(8), 11–16.

Schwandt, T. A. (2004). *"Sciencephobia" or legitimate worry? A diagnostic reading of science-based research*. Keynote address at the 2004 Minnesota Evaluation Studies Institute, Minneapolis.

Schwandt, T. A. (2005). Auditing. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 23–24). Thousand Oaks, CA: Sage.

Schweinhart, L. J., Barnes, H. V., & Weikart, D. P. (1993). *Significant benefits: The HighScope Perry Preschool study through age 27* (Monographs of the HighScope Educational Research Foundation #10). Ypsilanti, MI: HighScope Press.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39–83). Skokie, IL: Rand McNally.

Scriven, M. (1969a). *Evaluation skills* (Audiotape No. 6B). Washington, DC: American Educational Research Association.

Scriven, M. (1969b). An introduction to meta-evaluation. *Educational Products Report*, *2*(5), 36–38.

Scriven, M. (1973). Goal-free evaluation. In E. R. House (Ed.), *School evaluation: The politics and process* (pp. 319–328). Berkeley, CA: McCutchan.

Scriven, M. (1974). Pros and cons about goal-free evaluation. *Evaluation Comment*, *3*, 1–4.

Scriven, M. (1975). *Evaluation bias and its control* (Occasional Paper Series, Paper #4). Kalamazoo: Western Michigan University, Evaluation Center.

Scriven, M. (1980). *The logic of evaluation*. Inverness, CA: EdgePress.

Scriven, M. (1983). Evaluation ideologies. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and social services evaluation* (pp. 229–260). Norwell, MA: Kluwer.

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

Scriven, M. (1993). *Hard-won lessons in program evaluation*. New Directions for Program Evaluation, no. 58. San Francisco, CA: Jossey-Bass.

Scriven, M. (1994a). Evaluation as a discipline. *Studies in Educational Evaluation*, *20*, 147–166.

Scriven, M. (1994b). The final synthesis. *Evaluation Practice*, *15*, 367–382.

Scriven, M. (1994c). The fine line between evaluation and explanation. *Evaluation Practice*, *15*, 75–77.

Scriven, M. (1994d). Product evaluation: The state of the art. *Evaluation Practice*, *15*, 45–62.

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

Scriven, M. (1997). Empowerment evaluation examined. *Evaluation Practice*, *18*, 165–175.

Scriven, M. (1998). Minimalist theory: The least theory that practice requires. *American Journal of Evaluation*, *19*, 57–70.

Scriven, M. (2004a). The fiefdom problem. *Journal of MultiDisciplinary Evaluation*, *1*(1), 11–18.

Scriven, M. (2004b). Reflections. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 183–195). Thousand Oaks, CA: Sage.

Scriven, M. (2005a). *Can we infer causation from cross-sectional data?* Washington, DC: National Academy of Sciences.

Scriven, M. (2005b). Causation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 43–47). Thousand Oaks, CA: Sage.

Scriven, M. (2005c). Checklists. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 53–59). Thousand Oaks, CA: Sage.

Scriven, M. (2005d). Empowerment evaluation principles in practice [Review of the book *Empowerment evaluation principles in practice*, by D. M. Fetterman & A. Wandersman (Eds.). *American Journal of Evaluation*, *26*, 415–417.

Scriven, M. (2005e, September). *Theory-free evaluation*. Paper presented at the Evaluation Center's Evaluation Café, Western Michigan University, Kalamazoo.

Scriven, M. (2007). *Key Evaluation Checklist (KEC)*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/kec_feb07.pdf

Scriven, M. (2009a). Demythologizing causation and evidence. In S. I. Donaldson, C. A. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and program evaluation practice?* (pp. 134–152). Thousand Oaks, CA: Sage.

Scriven, M. (2009b). Meta-evaluation revisited. *Journal of MultiDisciplinary Evaluation*, *6*(11), iii–viii.

Scriven, M. (2011a). *Conceptual revolutions in evaluation: Past, present, and future*. Unpublished manuscript.

Scriven, M. (2011b). *Evaluating evaluations: A meta-evaluation checklist*. Claremont, CA: Claremont Graduate University.

Scriven, M. (2011c). Evaluation bias and its control*. *Journal of MultiDisciplinary Evaluation*, *7*(15), 79–98.

Scriven, M. (2011d). The Faster Forward Fund. *Journal of MultiDisciplinary Evaluation*, *7*(15), 313–317.

Scriven, M. (2011e). *The three revolutions*. Retrieved from http://michaelscriven.info/fasterforwardfund.html.

Scriven, M., & Coryn, C.L.S. (2008). The logic of research evaluation. In C.L.S. Coryn & M. Scriven (Eds.), *Reforming the evaluation of research* (pp. 89–106). New Directions for Evaluation, no. 118. San Francisco, CA: Jossey-Bass.

Scriven, M., & Roth, J. (1990). Special feature: Needs assessment. *Evaluation Practice*, *11*, 135–144.

Sechrest, L. E. (1997). Empowerment evaluation: Knowledge and tools for self-assessment and accountability. *Environment and Behavior*, *29*, 422–426.

Segone, N., & Ocampo, A. (Eds.). (2006). *Creating and developing evaluation organizations: Lessons learned from Africa, Americas, Asia, Australasia and Europe*. Lima, Peru: International Organisation for Cooperation in Evaluation.

Seidman, I. (2006). *Interviewing as qualitative research: A guide for researchers in education and the social sciences* (3rd ed.). New York, NY: Teachers College Press.

Seigart, D. (2005). Feminist evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 154–157). Thousand Oaks, CA: Sage.

Seigart, D., & Brisolara, S. (Eds.). (2002). *Feminist evaluation: Explorations and experiences*. New Directions for Evaluation, no. 96. San Francisco, CA: Jossey-Bass.

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*, 1–19.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Thousand Oaks, CA: Sage.

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. *Psychological Methods*, *16*, 179–191.

Shadish, W. R., & Luellen, J. K. (2005). History of evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 183–186). Thousand Oaks, CA: Sage.

Shadish, W. R., Newman, D. L., Scheirer, M. A., & Wye, C. (1995a). Developing the guiding principles. In W. R. Shadish, D. L. Newman, M. A. Scheirer, & C. Wye (Eds.), *Guiding principles for evaluators* (pp. 3–18). New Directions for Program Evaluation, no. 66. San Francisco, CA: Jossey-Bass.

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

Shenson, H. L. (1990). *The contract and fee-setting guide for consultants and professionals*. Hoboken, NJ: Wiley.

Shepard, L. A. (1977). *A checklist for evaluating large-scale assessment programs*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/wp-content/uploads/2010/05/assessment_eval.pdf

Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, *49*, 79–105.

Smith, E. R., & Tyler, R. W. (1942). *Appraising and recording student progress*. New York, NY: HarperCollins.

Smith, L. M., & Pohland, P. A. (1974). Educational technology and the rural highlands. In L. M. Smith (Ed.), *Four examples: Economic, anthropological, narrative, and portrayal* (pp. 13–52). Skokie, IL: Rand McNally.

Smith, M. F. (1986). The whole is greater: Combining qualitative and quantitative approaches in evaluation studies. In D. Williams (Ed.), *Naturalistic evaluation* (pp. 37–54). San Francisco, CA: Jossey-Bass.

Smith, M. F. (1989). *Evaluability assessment: A practical approach*. Norwell, MA: Kluwer.

Smith, N. L. (1987). Toward the justification of claims in evaluation research. *Evaluation and Program Planning*, *10*, 309–314.

Smith, N. L. (1993). Improving evaluation theory through the empirical study of evaluation practice. *Evaluation Practice*, *14*, 237–242.

Smith, N. L., Chircop, S., & Mukherjee, P. (2000). Considerations on the development of culturally relevant evaluation standards. In C. Russon (Ed.), *The program evaluation standards in international settings* (pp. 29–40). Kalamazoo: Western Michigan University, Evaluation Center.

Spybrook, J. K. (2008). Are power analyses reported with adequate detail? Evidence from the first wave of group randomized trials funded by the Institute of Education Sciences. *Journal of Research on Educational Effectiveness*, *1*, 215–235.

Spybrook, J. K., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis*, *31*(3), 298–318.

Stake, R. E. (n.d.). *Checklist for negotiating an agreement to evaluate an educational program*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/negotiating.pdf

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* (pp. 58–72). Columbus, OH: State Superintendent of Public Instruction.

Stake, R. E. (1971). *Measuring what learners learn*. Urbana: University of Illinois, Center for Instructional Research and Curriculum Evaluation.

Stake, R. E. (1974). *Nine approaches to educational evaluation*. Urbana: University of Illinois, Center for Instructional Research and Curriculum Evaluation.

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 Series, Paper #5). Kalamazoo: Western Michigan University, Evaluation Center.

Stake, R. E. (1976). A theoretical statement of responsive evaluation. *Studies in Educational Evaluation*, *2*, 19–22.

Stake, R. E. (1983). Program evaluation, particularly responsive evaluation. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and social services evaluation* (pp. 287–310). Norwell, MA: Kluwer.

Stake, R. E. (1986). *Quieting reform*. Urbana: University of Illinois Press.

Stake, R. E. (1988). Seeking sweet water. In R. M. Jaeger (Ed.), *Methods for research in education* (pp. 253–300). Washington, DC: American Educational Research Association.

Stake, R. E. (1994). Case studies. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 236–247). Thousand Oaks, CA: Sage.

Stake, R. E. (1995). *The art of case study research*. Thousand Oaks, CA: Sage.

Stake, R. E. (2003). Responsive evaluation. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 63–68). Norwell, MA: Kluwer.

Stake, R. E. (2004a). Stake and responsive evaluation. In M. C. Alkin (Ed.), *Evaluation roots: tracing theorists' views and influences* (pp. 203–217). Thousand Oaks, CA: Sage.

Stake, R. E. (2004b). *Standards-based and responsive evaluation*. Thousand Oaks, CA: Sage.

Stake, R. E. (2005). *Multiple case study analysis*. New York, NY: Guilford Press.

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

Stake, R. E. (2013). Responsive evaluation IV. In M. C. Alkin (Ed.), *Evaluation roots: A wider perspective of theorists' views and influences* (2nd ed.; pp. 189–197). Thousand Oaks, CA: Sage.

Stake, R. E., & Davis, R. (1999). Summary evaluation of Reader Focused Writing for the Veterans Benefits Administration. *American Journal of Evaluation*, *20*, 323–343.

Stake, R. E., Easley, J., & Anastasiou, K. (1978). *Case studies in science education*. Washington, DC: National Science Foundation, Directorate for Science Education, Office of Program Integration.

Stauffer, S. (1941). Notes on the case study and the unique case. *Sociometry*, *4*, 349–357.

Steinmetz, A. (1983). The discrepancy evaluation model. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and social services evaluation* (pp. 79–100). Norwell, MA: Kluwer.

Stevahn, L., King, J. A., Ghere, G., & Minnema, J. (2005). Establishing essential competencies for program evaluators. *American Journal of Evaluation*, *26*, 43–59.

Stingley, T. (2010, September). *Wine evaluation*. Paper presented at the Evaluation Center's Evaluation Café, Western Michigan University, Kalamazoo.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory of qualitative research*. Thousand Oaks, CA: Sage.

Strauss, A. (1987). *Qualitative analysis for social scientists*. Cambridge, UK: Cambridge University Press.

Stufflebeam, D. L. (1966a). A depth study of the evaluation requirement. *Theory into Practice*, *5*, 121–133.

Stufflebeam, D. L. (1966b, January). *Evaluation under Title I of the Elementary and Secondary Education Act of 1967*. Address delivered at the Title I Evaluation Conference sponsored by the Michigan State Department of Education, Lansing.

Stufflebeam, D. L. (1967). The use and abuse of evaluation in Title III. *Theory into Practice*, *6*, 126–133.

Stufflebeam, D. L. (1969). Evaluation as enlightenment for decision making. In A. Walcott (Ed.), *Improving educational assessment and an inventory of measures of affective behavior* (pp. 41–73). Washington, DC: Association for Supervision and Curriculum Development.

Stufflebeam, D. L. (1971a). The relevance of the CIPP evaluation model for educational accountability. *Journal of Research and Development in Education*, *5*(1), 19–25.

Stufflebeam, D. L. (1971b). The use of experimental design in educational evaluation. *Journal of Educational Measurement*, *8*, 267–274.

Stufflebeam, D. L. (1974). *Meta-evaluation* (Occasional Paper Series, Paper #3). Kalamazoo: Western Michigan University, Evaluation Center.

Stufflebeam, D. L. (1978). Metaevaluation: An overview. *Evaluation & the Health Professions*, *1*(2), 146–163.

Stufflebeam, D. L. (1983). The CIPP model for program evaluation. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 117–141). Norwell, MA: Kluwer.

Stufflebeam, D. L. (1985). Stufflebeam's improvement-oriented evaluation. In D. L. Stufflebeam & A. J. Shinkfield, *Systematic evaluation: A self-instructional guide to theory and practice* (pp. 151–207). Norwell, MA: Kluwer.

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. (1997). A standards-based perspective on evaluation. In R. L. Stake (Ed.), *Evaluation and the postmodern dilemma* (Vol. 3, pp. 61–88). Greenwich, CT: Jai Press.

Stufflebeam, D. L. (1999a). *Evaluation Contracts Checklist*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/contracts.pdf

Stufflebeam, D. L. (1999b). *Program Evaluations Metaevaluation Checklist*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/program_metaeval_10point.pdf

Stufflebeam, D. L. (2000a). Lessons in contracting for evaluations. *American Journal of Evaluation*, *21*, 293–314.

Stufflebeam, D. L. (2000b). The methodology of metaevaluation. In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed., pp. 457–496). Norwell, MA: Kluwer.

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, CA: Jossey-Bass.

Stufflebeam, D. L. (2001c). The metaevaluation imperative. *American Journal of Evaluation*, *22*, 183–209.

Stufflebeam, D. L. (2003a). The CIPP model for evaluation. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 31–62). Norwell, MA: Kluwer.

Stufflebeam, D. L. (2003b). Institutionalizing evaluation in schools. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 775–806). Norwell, MA: Kluwer.

Stufflebeam, D. L. (2004a). *Evaluation Design Checklist*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/evaldesign.pdf

Stufflebeam, D. L. (2004b). The 21st century CIPP model: Origins, development, and use. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 245–266). Thousand Oaks, CA: Sage.

Stufflebeam, D. L. (2005). CIPP model (context, input, process, product). In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 60–65). Thousand Oaks, CA: Sage.

Stufflebeam, D. L. (2007). *CIPP Evaluation Model Checklist*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/cippchecklist_mar07.pdf

Stufflebeam, D. L. (2011a). Meta-evaluation. *Journal of MultiDisciplinary Evaluation*, *7*(15), 99–158.

Stufflebeam, D. L. (2011b). *Program Evaluations Metaevaluation Checklist*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/checklists

Stufflebeam, D. L. (2013). The CIPP evaluation model: Status, origins, development, use, and theory. In M. C. Alkin (Ed.), *Evaluation roots: A wider perspective of theorists' views and influences* (2nd ed.; pp. 243–260). Thousand Oaks, CA: Sage.

Stufflebeam, D. L., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond, R. L., Merriman, H. O., & Provus, M. (1971). *Educational evaluation and decision making*. Itasca, IL: Peacock.

Stufflebeam, D. L., Goodyear, L., Marquart, J., & Johnson, E. (2005). *Guiding principles checklist*. Kalamazoo: Western Michigan University, Evaluation Center. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/guidingprinciples2005.pdf

Stufflebeam, D. L., Gullickson, A., & Wingate, L. (2002). *The Spirit of Consuelo: An evaluation of Ke Aka Ho'ona*. Kalamazoo: Western Michigan University, Evaluation Center.

Stufflebeam, D. L., Jaeger, R. M., & Scriven, M. (1992, April). *A retrospective analysis of a summative evaluation of NAGB's pilot project to set achievement levels on the National Assessment of Educational Progress*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Stufflebeam, D. L., Madaus, G. F., & Kellaghan, T. (Eds.). (2000). *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed.). Norwell, MA: Kluwer.

Stufflebeam, D. L., McCormick, C. H., Brinkerhoff, R. O., & Nelson, C. O. (1985). *Conducting educational needs assessment*. Norwell, MA: Kluwer.

Stufflebeam, D. L., & Shinkfield, A. J. (1985). *Systematic evaluation: A self-instructional guide to theory and practice*. Norwell, MA: Kluwer.

Stufflebeam, D. L., & Shinkfield, A. J. (2007). *Evaluation theory, models, and applications*. San Francisco, CA: Jossey-Bass.

Stufflebeam, D. L., & Webster, W. J. (1988). Evaluation as an administrative function. In N. Boyan (Ed.), *Handbook of research on educational administration* (pp. 569–601). White Plains, NY: Longman.

Suchman, E. A. (1967). *Evaluative research: Principles and practice in public service and social action programs*. New York, NY: Russell Sage Foundation.

Sumida, J. (1994). *The Waianae self-help housing initiative: Ke Aka Ho'ona; Traveling observer handbook*. Kalamazoo: Western Michigan University, Evaluation Center.

Tarsilla, M. (2010a). Being blind in a world of multiple perspectives: The evaluator's dilemma between the hope of becoming a team player and the fear of becoming a critical friend with no friends. *Journal of MultiDisciplinary Evaluation*, *6*(13), 200–205.

Tarsilla, M. (2010b). Inclusiveness and social justice in evaluation: Can the transformative agenda really alter the status quo? A conversation with Donna M. Mertens. *Journal of MultiDisciplinary Evaluation*, *6*(14), 102–113.

Tashakkori, A., & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social and behavioral research*. Thousand Oaks, CA: Sage.

Taut, S. (2000). Cross-cultural transferability of the program evaluation standards. In C. Russon (Ed.), *The program evaluation standards in international settings* (pp. 5–28). Kalamazoo: Western Michigan University, Evaluation Center.

Tesch, R. (1990). *Qualitative research: Analysis types and software tools*. Bristol, PA: Falmer.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York, NY: Guilford Press.

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, CA: Jossey-Bass.

Torres, R. T. (1991). Improving the quality of internal evaluation: The evaluator as consultant mediator. *Evaluation and Program Planning*, *14*, 189–198.

Torres, R. T., Preskill, H., & Piontek, M. E. (2005). *Evaluation strategies for communicating and reporting: Enhancing learning in organizations* (2nd ed.). Thousand Oaks, CA: Sage.

Travers, R.M.W. (1983). *How research has changed American schools*. Kalamazoo, MI: Mythos Press.

Trochim, W.M.K. (1984). *Research design for program evaluation: The regression discontinuity approach*. Thousand Oaks, CA: Sage.

Trochim, W.M.K., & Cappelleri, J. C. (1992). Cutoff assignment strategies for enhancing randomized clinical trials. *Controlled Clinical Trials*, *13*, 190–212.

Tsang, M. C. (1997). Cost analysis for improved educational policymaking and evaluation. *Educational Evaluation and Policy Analysis*, *19*(4), 318–324.

Tufte, E. R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (1997). *Visual explanations: Images and quantities, evidence and narrative*. Cheshire, CT: Graphics Press.

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Tyler, R. W. (1932). *Service studies in higher education*. Columbus: Ohio State University, Bureau of Educational Research.

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, IL: University of Chicago Press.

Tyler, R. W. (1966). The objectives and plans for a national assessment of educational progress. *Journal of Educational Measurement*, *3*, 1–10.

Tyler, R. W. (1967). Changing concepts of educational evaluation. In R. E. Stake (Ed.), *Perspectives of curriculum evaluation* (pp. 13–18). Skokie, IL: Rand McNally.

Tyler, T. R., Boeckmann, R. J., Smith, H. J., & Huo, Y. J. (1997). *Social justice in a diverse society*. Oxford, UK: Westview Press.

Tymms, P. (1995). *Setting up a national "value-added" system for primary education in England: Problems and possibilities*. Paper presented at the National Evaluation Institute, Kalamazoo, MI.

U.S. Department of Education. (2003, November 4). Scientifically based evaluation methods. *Federal Register*, *68*(213). RIN 1890-ZA00. Retrieved from http://www.eval.org/doesstatement.htm

U.S. General Accounting Office. (2002). *Government auditing standards: Amendment no. 3 Independence* (GAO-02-388G). Washington, DC: U.S. Government Printing Office.

U.S. Government Accountability Office. (2003). *Government auditing standards* (GAO-03-763G). Washington, DC: U.S. Government Printing Office.

U.S. Government Accountability Office. (2007). *Government auditing standards* (GAO-07-731G). Washington, DC: U.S. Government Printing Office.

U.S. Government Accountability Office. (2009). *A variety of rigorous methods can help identify effective interventions* (GAO-10–30). Washington, DC: U.S. Government Printing Office.

Vallance, E. (1973). *Aesthetic criticism and curriculum description*. Unpublished doctoral dissertation, Stanford University, Stanford, CA.

Vinovskis, M. (1999). *Overseeing the nation's report card: The creation and evolution of the National Assessment Governing Board (NAGB)*. Washington, DC: National Assessment Governing Board.

Walton, M. (1986). *The Deming management method*. New York, NY: Putnam.

Wandersman, A., Snell-Johns, J., Lentz, B. E., Fetterman, D. M., Keener, D. C., Livet, M., . . . Flaspohler, P. (2005). The principles of empowerment evaluation. In D. Fetterman & A. Wandersman (Eds.), *Empowerment evaluation principles in practice* (pp. 27–41). New York, NY: Guilford Press.

Weaver, L., & Cousins, J. B. (2004). Unpacking the participatory process. *Journal of MultiDisciplinary Evaluation*, *1*(1), 19–40.

Webster, W. J. (1975, March). *The organization and functions of research evaluation in a large urban school district*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Webster, W. J. (1995). The connection between personnel evaluation and school evaluation. *Studies in Educational Evaluation*, *21*, 227–254.

Webster, W. J., Mendro, R. L., & Almaguer, T. O. (1994). Effectiveness indices: A "value-added" approach to measuring school effect. *Studies in Educational Evaluation*, *20*, 113–145.

Weiss, C. H. (1972). *Evaluation*. Englewood Cliffs, NJ: Prentice Hall.

Weiss, C. H. (1983). The stakeholder approach to evaluation: Origins and promise. In A. S. Bryk (Ed.), *Stakeholder-based evaluation* (pp. 3–14). New Directions for Program Evaluation, no. 17. San Francisco, CA: Jossey-Bass.

Weiss, C. H. (1995). Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In J. Connell, A. Kubisch, L. B. Schorr, & C. H. Weiss (Eds.), *New approaches to evaluating community initiatives: Vol. 1. Concepts, methods, and contexts* (pp. 65–92). New York, NY: Aspen Institute

Weiss, C. H. (1997a). How can theory-based evaluations make greater headway? *Evaluation Review*, *21*, 501–524.

Weiss, C. H. (1997b). Theory-based evaluation: Past, present and future. In D. J. Rog & D. Fournier (Eds.), *Progress and future directions in evaluation: Perspectives on theory, practice and methods* (pp. 41–55). New Directions for Evaluation, no. 76. San Francisco, CA: Jossey-Bass.

Weiss, C. H. (1998). *Evaluation: Methods for studying programs and policies* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Weiss, C. H. (2000). Which links in which theories shall we evaluate? In P. J. Rogers, T. A. Hasci, A. Petrosino, & T. A. Huebner (Eds.), *Program theory in evaluation: Challenges and opportunities* (pp. 35–45). New Directions for Evaluation, no. 87. San Francisco, CA: Jossey-Bass.

Weiss, C. H. (2004a). On theory-based evaluation: Winning friends and influencing people. *The Evaluation Exchange*, *9*(5), 1–5.

Weiss, C. H. (2004b). Rooting for evaluation: A Cliff Notes version of my work. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 153–168). Thousand Oaks, CA: Sage.

Whitmore, E. (Ed.). (1998). *Understanding and practicing participatory evaluation*. New Directions for Evaluation, no. 80. San Francisco, CA: Jossey-Bass.

Wholey, J. S. (1995). Assessing the feasibility and likely usefulness of evaluation. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (pp. 15–39). San Francisco, CA: Jossey-Bass.

Wholey, J. S. (1996). Formative and summative evaluation: Related issues in performance measurement. *Evaluation Practice*, *17*, 145–149.

Widmer, T., Landert, C., & Bacmann, N. (2000). Evaluation standards recommended by the Swiss Evaluation Society (SEVAL). In C. Russon (Ed.), *The program evaluation standards in international settings* (pp. 81–102). Kalamazoo: Western Michigan University, Evaluation Center.

Wiersma, W., & Jurs, S. G. (2005). *Research methods in education: An introduction* (8th ed.). Needham Heights, MA: Allyn & Bacon.

Wiersma, W., & Jurs, S. G. (2009). *Research methods in education: An introduction* (9th ed.). Boston, MA: Pearson.

Wilkinson, L., & Taskforce on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 595–604.

Williams, B., & Imam, I. (Eds.). (2007). *Systems concepts in evaluation: An expert anthology*. Inverness, CA: EdgePress.

Winer, B. J. (1962). *Statistical principles in experimental design*. New York, NY: McGraw-Hill.

Wingate, L. (2009). *The program evaluation standards applied for meta-evaluation purposes: Investigating interrater reliability and its implications for use*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Wisler, C. (Ed.). (1996). *Evaluation and auditing: Prospects for convergence*. New Directions for Program Evaluation, no. 71. San Francisco, CA: Jossey-Bass.

Wolcott, H. F. (1995). *The art of fieldwork*. Walnut Creek, CA: AltaMira.

Wolf, R. L. (1975). Trial by jury: A new evaluation method. *Phi Delta Kappan*, *3*(57), 185–187.

Worthen, B. R. (1999). Critical challenges confronting certification of evaluators. *American Journal of Evaluation*, *20*, 533–555.

Worthen, B. R., & Sanders, J. R. (1987). *Educational evaluation: Alternative approaches and practical guidelines*. White Plains, NY: Longman.

Worthen, B. R., Sanders, J. R., & Fitzpatrick, J. L. (1997). *Program evaluation: Alternative approaches and practical guidelines* (2nd ed.). New York, NY: Longman.

Yates, B. T. (1996). *Analyzing costs, procedures, processes, and outcomes in human services*. Applied Social Research Methods Series, Vol. 42. Thousand Oaks, CA: Sage.

Yau, N. (2011). *Visualize this: The FlowingData guide to design, visualization, and statistics*. Hoboken, NJ: Wiley.

Yin, R. K. (1991). *Case study research: Design and methods*. Thousand Oaks, CA: Sage.

Yin, R. K. (1992). The case study as a tool for doing evaluation. *Current Sociology*, *40*(1), 121–137.

Yin, R. K. (1998). The abridged version of case study research: Design and method. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 229–260). Thousand Oaks, CA: Sage.

Yin, R. K. (2009). *Case study research: Design and methods* (4th ed.). Applied Social Research Methods Series, Vol. 5. Thousand Oaks, CA: Sage.

Zhang, G., Zeller, N., Griffith, R., Metcalf, D., Shea, C., Williams, J., & Misulis, K. (2010, October). *Using the CIPP model as a comprehensive framework to guide the planning, implementation, and assessment of service-learning programs*. Paper presented at the annual meeting of the National Evaluation Institute, Williamsburg, VA.

INDEX

B

C

D

E

F

G

H

J

K

L

M

N

Q

R

S

T

U

V

W

Y

Z

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.