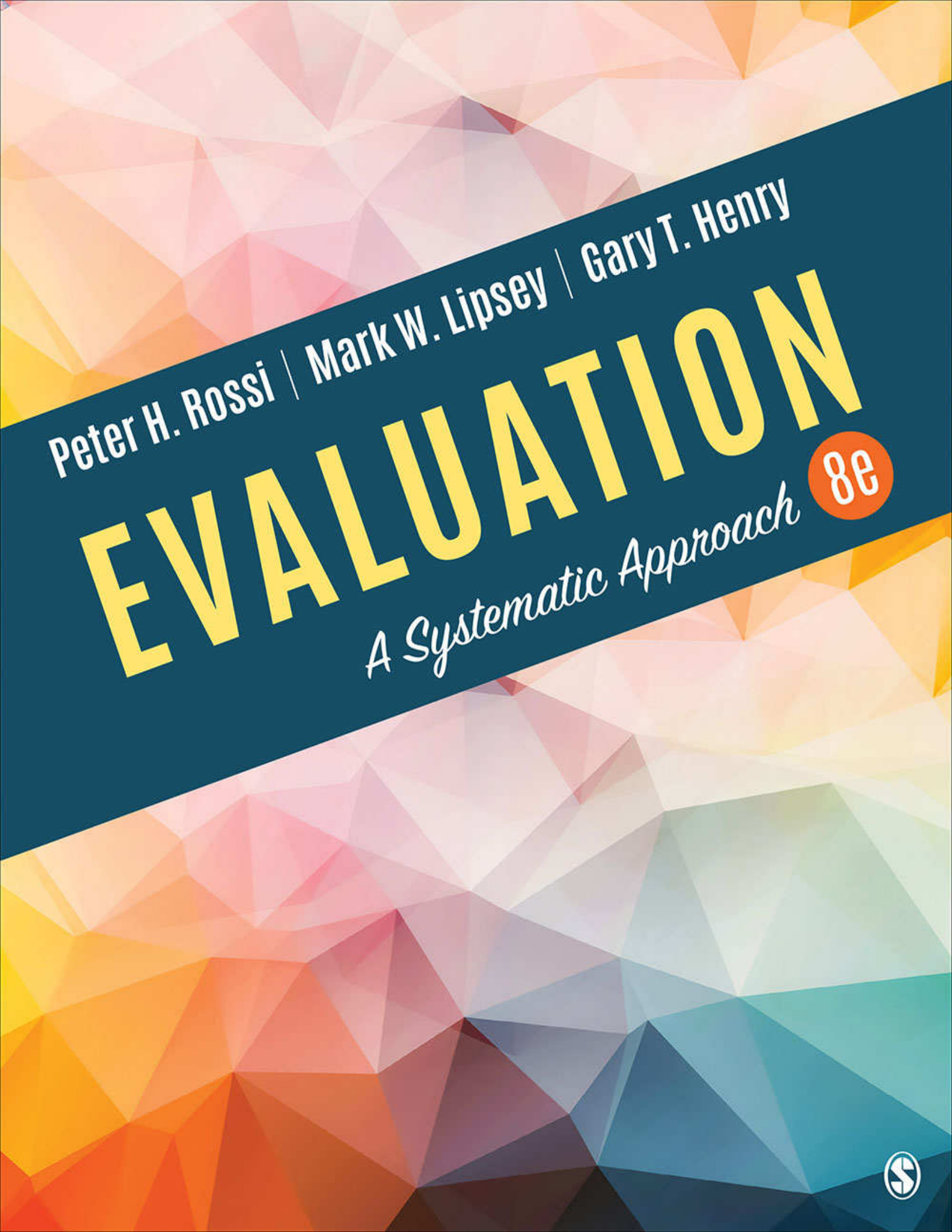




Peter H. Rossi | Mark W. Lipsey | Gary T. Henry

EVALUATION

A Systematic Approach

8e



Praise for the Eighth Edition

“This thoughtfully designed text provides an update on a classic tool for understanding evaluation.”

—*Brian Boggs, University of Michigan and Michigan State University*

“The eighth edition continues to offer broad instruction in program evaluation concepts, methods, and practice, from planning to communicating results. The addition of critical thinking and discussion questions provides the opportunity for classroom discussion as well as application of concepts. I recommend this text for use with master’s and doctoral level students.”

—*Nancy Bridier, Grand Canyon University*

“The eighth edition is a wonderful resource for professional degree students, and can also provide a practical component for students taking a practicum class.”

—*Raven Brown, Baruch College, CUNY*

“An excellent and concise book defining the systematic approach to program evaluation: the best resource for both students and researchers.”

—*Anil Kumar Chaudhary, Pennsylvania State University*

“The long-awaited eighth edition includes materials and chapters that reflect current developments in the field of evaluation research. The new edition, with substantive revisions, provides foundational knowledge and perspectives on evaluation without losing the legacy and wisdom of Dr. Rossi.”

—*Young Ik Cho, University of Wisconsin–Milwaukee*

“The eighth edition is a massive improvement on an already stellar text. The breadth and depth of coverage, while still keeping a practical focus, make this the go-to book for program evaluation classes and practitioners alike.”

—*B. Andrew Chupp, Indiana University*

“As a professor and a program evaluator, I find that this book presents a realistic, pragmatic view of program evaluation. Clearly presented, the authors use the same language I use with clients, which helps to ease students’ transition to the workplace.”

—*Leslie Eaton, SUNY Cortland*

“This book truly represents the gold standard on everything one would want or need to know about program evaluation, including checklists and diagrams. The Planning an Evaluation chapter basically provides a step-by-step guide to performing a program evaluation with as much rigor as possible. The entire text is rich with examples of actual program assessments.”

—*Kristin Grosskopf, University of Nebraska–Lincoln*

“An earlier version of this text was useful to me as an evaluation student. This revised version will ensure that today’s students have an invaluable resource that clearly communicates what is unique about our field, while also introducing the range of approaches and methods that evaluators may use.”

—*Melissa Haynes, University of Minnesota*

“The material in the eighth edition is effectively sequenced, and the technical orientation of the chapters makes the book an indispensable partner even for seasoned scholars and practitioners in the art of program evaluation.”

—*Kalu Kalu, Auburn University at Montgomery*

“This book offers a comprehensive view of evaluation and serves as a valuable guide in developing an evaluation plan.”

—*Sarmithsa Majumdar, Texas Southern University*

“The authors do a phenomenal job of unpacking complex terms and ideas making this reading accessible to learners.”

—*Jessica Wendorf Muhamad, Florida State University*

“This is another exceptional work by the authors. This book not only helps novice evaluators, it also provides tools for expert evaluators. This new edition brings new cases and exhibits that connect the theory to practice and contextualizes the content for students.”

—*Osman Özturgut, California State University Channel Islands*

“The eighth edition covers the essentials of evaluation extremely well, serves as a guide for development of specific approaches of evaluation, and enhances the critical thinking of students.”

—*David Pugh, Edinboro University*

“I have used previous editions of this text either as a student or professor for 20 years. This new edition is a great update of a reliable textbook on evaluation, including updated terminology and methodology.”

—*Kimberley Shoaf, University of Utah*

Sara Miller McCune founded SAGE Publishing in 1965 to support the dissemination of usable knowledge and educate a global community. SAGE publishes more than 1000 journals and over 800 new books each year, spanning a wide range of subject areas. Our growing selection of library products includes archives, data, case studies and video. SAGE remains majority owned by our founder and after her lifetime will become owned by a charitable trust that secures the company's continued independence.

Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

Evaluation

A Systematic Approach

Eighth Edition

Peter H. Rossi
Mark W. Lipsey
Vanderbilt University
Gary T. Henry
Vanderbilt University



Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

Los Angeles
London
New Delhi
Singapore
Washington DC
Melbourne



FOR INFORMATION:

SAGE Publications, Inc.

2455 Teller Road

Thousand Oaks, California 91320

E-mail: order@sagepub.com

SAGE Publications Ltd.

1 Oliver's Yard

55 City Road

London EC1Y 1SP

United Kingdom

SAGE Publications India Pvt. Ltd.

B 1/I 1 Mohan Cooperative Industrial Area

Mathura Road, New Delhi 110 044

India

SAGE Publications Asia-Pacific Pte. Ltd.

18 Cross Street #10-10/11/12

China Square Central

Singapore 048423

Copyright © 2019 by SAGE Publications, Inc.

All rights reserved. Except as permitted by U.S. copyright law, no part of this work may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without permission in writing from the publisher.

All third-party trademarks referenced or depicted herein are included solely for the purpose of illustration and are the property of their respective owners. Reference to these trademarks in no way indicates any relationship with, or endorsement by, the trademark owner.

Printed in the United States of America

ISBN 978-1-5063-0788-6

This book is printed on acid-free paper.

18 19 20 21 22 10 9 8 7 6 5 4 3 2 1

Acquisitions Editor: Helen Salmon

Content Development Editor: Chelsea Neve

Editorial Assistant: Megan O’Heffernan

Production Editor: Olivia Weber-Stenis

Copy Editor: Jim Kelly

Typesetter: C&M Digital (P) Ltd.

Proofreader: Victoria Reed-Castro

Indexer: Maria Sosnowski

Cover Designer: Candice Harman

Marketing Manager: Susannah Coldes

Brief Contents Chapter Outline Chapter Outline Chapter Outline Chapter Outline Chapter Outline Chapter Outline Chapter Outline Chapter Outline Chapter Outline Chapter Outline Chapter Outline Chapter Outline

[Preface](#)

[Acknowledgments](#)

[About the Authors](#)

[1 | What Is Program Evaluation and Why Is It Needed?](#)

[2 | Social Problems and Assessing the Need for a Program](#)

[3 | Assessing Program Theory and Design](#)

[4 | Assessing Program Process and Implementation](#)

[5 | Measuring and Monitoring Program Outcomes](#)

[6 | Impact Evaluation: Isolating the Effects of Social Programs in the Real World](#)

[7 | Impact Evaluation: Comparison Group Designs](#)

[8 | Impact Evaluation: Designs With Strict Controls on Program Access](#)

[9 | Detecting, Interpreting, and Exploring Program Effects](#)

[10 | Assessing the Economic Efficiency of Programs](#)

[11 | Planning an Evaluation](#)

[12 | The Social and Political Context of Evaluation](#)

[Glossary](#)

[References](#)

[Author Index](#)

[Subject Index](#)

Detailed Contents

[Preface](#)

[Acknowledgments](#)

[About the Authors](#)

[1 | What Is Program Evaluation and Why Is It Needed?](#)

[What Is Program Evaluation?](#)

[Why Is Program Evaluation Needed?](#)

[Systematic Program Evaluation](#)

[The Central Role of Evaluation Questions](#)

[The Five Domains of Evaluation Questions and Methods](#)

[Summary](#)

[Key Concepts](#)

[2 | Social Problems and Assessing the Need for a Program](#)

[The Role of Evaluators in Diagnosing Social Conditions and Service Needs](#)

[Defining the Problem to Be Addressed](#)

[Specifying the Extent of the Problem: When, Where, and How Big?](#)

[Defining and Identifying the Target Populations of Interventions](#)

[Describing Target Populations](#)

[Describing the Nature of Service Needs](#)

[Summary](#)

[Key Concepts](#)

[3 | Assessing Program Theory and Design](#)

[Evaluability Assessment](#)

[Describing Program Theory](#)

[Eliciting Program Theory](#)

[Assessing Program Theory](#)

[Possible Outcomes of Program Theory Assessment](#)

[Summary](#)

[Key Concepts](#)

[4 | Assessing Program Process and Implementation](#)

[What Is Program Process Evaluation and Monitoring?](#)

[Perspectives on Program Process Monitoring](#)

[Assessing Service Utilization](#)

[Assessing Organizational Functions](#)

[Summary.](#)

[Key Concepts](#)

[5 | Measuring and Monitoring Program Outcomes](#)

[Program Outcomes](#)

[Identifying Relevant Outcomes](#)

[Measuring Program Outcomes](#)

[Monitoring Program Outcomes](#)

[Summary.](#)

[Key Concepts](#)

[6 | Impact Evaluation: Isolating the Effects of Social Programs in the Real World](#)

[The Nature and Importance of Impact Evaluation](#)

[When Is an Impact Evaluation Appropriate?](#)

[What Would Have Happened Without the Program?](#)

[The Logic of Impact Evaluation: The Potential Outcomes](#)

[Framework](#)

[The Fundamental Problem of Causal Inference: Unavoidable](#)

[Missing Data](#)

[Summary.](#)

[Key Concepts](#)

[7 | Impact Evaluation: Comparison Group Designs](#)

[Bias in Estimation of Program Effects](#)

[Potential Advantages of Comparison Group Designs](#)

[Comparison Group Designs for Impact Evaluation](#)

[Cautions About Quasi-Experiments for Impact Evaluation](#)

[Summary.](#)

[Key Concepts](#)

[8 | Impact Evaluation: Designs With Strict Controls on Program Access](#)

[Controlling Selection Bias by Controlling Access to the Program](#)

[Key Concepts in Impact Evaluation](#)

[When Is Random Assignment Ethical and Practical?](#)

[Application of the Regression Discontinuity Design](#)

[Choosing an Impact Evaluation Design](#)

[Summary.](#)

[Key Concepts](#)

[9 | Detecting, Interpreting, and Exploring Program Effects](#)

[The Magnitude of a Program Effect](#)
[Detecting Program Effects](#)
[Examining Variation in Program Effects](#)
[The Role of Meta-Analysis](#)
[Summary](#)
[Key Concepts](#)

10 | [Assessing the Economic Efficiency of Programs](#)
[Key Concepts in Efficiency Analysis](#)
[Conducting Cost-Benefit Analyses](#)
[Conducting Cost-Effectiveness Analyses](#)
[Summary](#)
[Key Concepts](#)

11 | [Planning an Evaluation](#)
[Evaluation Purpose and Scope](#)
[Data Collection, Acquisition, and Management](#)
[Data Analysis Plan](#)
[Communication Plan](#)
[Project Management Plan](#)
[Summary](#)
[Key Concepts](#)

12 | [The Social and Political Context of Evaluation](#)
[The Social Ecology of Evaluations](#)
[The Profession of Evaluation](#)
[Evaluation Standards, Guidelines, and Ethics](#)
[Utilization of Evaluation Results](#)
[Epilogue: The Future of Evaluation](#)
[Summary](#)
[Key Concepts](#)

[Glossary](#)

[References](#)

[Author Index](#)

[Subject Index](#)

*To the memory of Peter H. Rossi—
intellectual, scholar, policy researcher, colleague,
and program evaluation trailblazer*

Preface

Program evaluation is relatively new as a recognized area of organized activity. It was only in the 1970s that the first journals with *evaluation* in the title were launched, the first professional organizations were formed, and the first textbooks were published. One of the earliest of those textbooks was the first edition of this one, *Evaluation: A Systematic Approach*, authored by Peter Rossi, Howard Freeman, and Sonia Rosenbaum and published in 1979. With the benefit of hindsight, it is easy to recognize what a landmark that was. The publication of that comprehensive text at that time marked the point at which program evaluation had come of age as a field of endeavor with its own distinct identity, concepts, methods, and practices.

From 1982 through 1993, Rossi and Freeman updated this classic text with successive editions until, after the fifth edition, Mark Lipsey joined as a coauthor and helped produce the sixth and seventh editions. With this long history, *Evaluation: A Systematic Approach* has not only mirrored the evolution of program evaluation as a field of study, but helped shape that evolution. Peter Rossi and Howard Freeman, whose perspective on evaluation is now an indelible part of this history, are no longer with us. However, their contributions live on in this eighth edition, which we are proud to introduce in the spirit of the periodic updating and refreshing of this text that is part of their legacy.

And it is in that same spirit that Gary Henry has come on board as the newest coauthor, bringing energy, insight, and wisdom to the revisions embodied in this new edition. Gary has a wealth of practical evaluation experience to draw on and a deep understanding of the concepts, methods, and history of the field, all of which has helped bring this eighth edition to its current full development. While Lipsey and Henry take responsibility for the contents of this newest edition, Peter Rossi's hand is evident in much of the structure, orientation, and philosophy of this volume, and we honor that continuity by recognizing him as the lead author of this enduring text.

What has not changed is the intended audience for this textbook. It is written to introduce master's- and doctoral-level students to the concepts, methods, and practice of contemporary program evaluation research, and serves as well for those in professional positions involving evaluation who have not had the opportunity to be exposed to such an introduction. As such, this textbook provides an overview of all the major domains of evaluation: needs assessment, program theory, process evaluation, impact evaluation, and cost-effectiveness. Moreover, as in previous editions, these evaluation domains are presented in a coherent framework that not only explores each but recognizes their interrelationships, their role in improving social programs and the outcomes they are designed to affect, and their embeddedness in social and political context.

Furthermore, because of the varied program areas in which evaluators work, the coverage of these topics spans a range of application areas, including, public policy, education, welfare, criminal justice, public health, behavioral health, social work, and the like. This book can therefore be used as the primary text for graduate-level courses in any of these disciplines with only modest supplementary readings selected by the instructor to highlight issues and applications in the respective discipline. The common theme across these application areas is an applied research perspective that prioritizes credible evidence in support of informed and effective practice and policy.

New to This Edition

While maintaining continuity with the general structure and orientation of prior editions, this eighth edition incorporates a number of new or enhanced features and some substantial revisions. The most noteworthy of these are the following:

- A revised introductory chapter that condenses some of the topics spread across the first three chapters of the prior edition to provide a more efficient introduction with an emphasis on the distinctive characteristics of systematic program evaluation research and why it is essential for guiding effective policy and practice.
- Expanded coverage of the concepts and methods of impact evaluation in light of its relevance to the increased attention during the past decade given to developing, identifying, and implementing evidence-based programs and practices. This expanded coverage includes more in-depth treatment of evaluation designs that have become more prevalent in recent years, such as regression discontinuity and interrupted time series, and uses the increasingly relevant potential outcomes framework for explaining contemporary thinking about impact evaluation design and implementation. In addition, we discuss methods for developing more nuanced perspectives on program effects via analysis of moderators, mediators, and variation in implementation fidelity.
- A new chapter that provides practical guidance for planning an evaluation with a full discussion of the various components of an evaluation plan from expressing its purpose and design to negotiating intellectual property rights to planning the communication of the findings to achieve influence.
- A set of critical thinking and discussion questions plus a set of suggested application exercises for students at the end of each chapter. These are designed to assist instructors who wish to facilitate the active engagement of students with the issues and concepts covered in each chapter.
- Updates and revisions to every chapter to refresh the content and coverage and include current exhibits and examples drawn from a

wide range of program and policy areas. These include examples from large-scale evaluations and smaller local evaluations as well as examples from every corner of the globe.

We believe these updates, revisions, and new features make this classic text more engaging, informative, and current with the state of the art in program evaluation. We would be very pleased to receive feedback and suggestions for further improvements that could be made in future editions from instructors and students who use this book (mark.lipsey@vanderbilt.edu; gary.henry@vanderbilt.edu).

Companion Website

Evaluation: A Systematic Approach, Eighth Edition, is accompanied by a companion website featuring an array of free learning and teaching tools for both students and instructors.

The companion website is available at <https://study.sagepub.com/rossi8e>.

Password-protected **Instructor Resources** include:

- Editable, chapter-specific **PowerPoint® slides** that offer flexibility when creating multimedia lectures. Slides can be customized to meet your exact needs.
- **Essay questions** that assess students' understanding and application of the concepts. Questions can be given as homework or exams, and suggested answers are included to facilitate grading.
- **Tables and figures from the book** available for download and use in your course.

Open-access **Student Resources** include:

- Carefully selected **SAGE journal articles** illustrate the concepts presented in each chapter. Toll-free links provide direct access for readers.

Acknowledgments

It doesn't quite take a village to revise a textbook with the established history of this one, but it does take a team, and some members of that team labor behind the scenes and deserve a public thanks for their contributions. The Sage Publications crew that has turned our word-processed manuscripts into an actual book certainly fall into this category. A special thanks also goes to our editor, Helen Salmon, for her patience with our lapses and her gentle nudges about the importance of staying on schedule despite our frequent violations of said schedule.

We are also greatly appreciative of the contributions of Amy Donley, who has drafted the discussion questions and application exercises that appear at the end of each chapter. Amy is an assistant professor in sociology and director of the Institute for Social and Behavioral Sciences at the University of Central Florida who has taught evaluation courses using the previous edition of this textbook. That experience and her insights about how to engage students in the challenges of evaluation research have been invaluable for effectively focusing the content of these pedagogical aids.

And, among those who have toiled backstage, we need to bring to center stage for a bow and a round of applause the two graduate research assistants in the Department of Leadership, Policy, and Organizations at Vanderbilt—Catherine Kelly and Maryia Krivouchko—who have devoted countless hours to searching the evaluation and policy literature for timely and appropriate examples of key concepts, proofreading, compiling references, and organizing glossary terms and definitions. For them, we offer our own standing ovation.

About the Authors

Peter H. Rossi

was the lead author of the first edition of *Evaluation: A Systematic Approach* (1979) and of every successive edition through the seventh, published in 2004. His death in 2006 was a loss in more ways than can be enumerated, one of which was his engaged role in this textbook series. The punctuation at that point in this series might rightfully have been a period, and the seventh could have been the last edition. However, Peter would not have wanted that if the orientation and philosophy inherent in all the volumes of this series could be continued. With that in mind, the current coauthors have endeavored to produce an eighth edition that keeps that orientation and philosophy intact and thus recognizes Peter's continuing influence as the guiding hand that has shaped the results.

Even without the enduring contributions the *Evaluation* textbook series has made to the field of program evaluation, Peter Rossi stands tall among the small group of trailblazers whose vision and exemplary evaluation studies gave name and life to the emerging field of program evaluation in the 1970s. He had the stature of someone who had served on the faculties of such distinguished universities as Harvard, the University of Chicago, Johns Hopkins, and the University of Massachusetts at Amherst, where he finished his career as the Stuart A. Rice Professor Emeritus of Sociology. He had extensive applied research experience, including terms as the director of the National Opinion Research Center and director of the Social and Demographic Research Institute at the University of Massachusetts at Amherst. He conducted landmark studies that were models of high-quality, policy-relevant research in such areas as welfare reform, poverty, homelessness, criminal justice, and family preservation and wrote prolifically about them. And he also wrote about the theory, methods, and practice of program evaluation in ways that helped give shape to this new field of endeavor. The orientation and philosophy Peter brought to this work, and that endures in the new eighth edition of *Evaluation*, is a belief that, above all, the facts matter and thus are the

proper basis for any contribution of program evaluators to policy or practice. With respect for the facts comes respect for the methods that best elucidate those facts, and Peter was an unstinting champion of using the strongest feasible methods to tackle questions about social programs.

Mark W. Lipsey

recently stepped down as the director of the Peabody Research Institute at Vanderbilt University, a research unit devoted to research on interventions for at-risk populations. After a more than 40-year career in program evaluation, he has recently transitioned to what he calls “semiretirement” but maintains an appointment as a research professor in the Peabody College Department of Human and Organizational Development. His research specialties are evaluation research and research synthesis (meta-analysis) investigating the effects of social interventions with children, youth, and families. The topics of his recent work have been risk and intervention for juvenile delinquency and substance use, early childhood education programs, issues of methodological quality in program evaluation, and ways to help practitioners and policymakers make better use of research to improve the outcomes of programs for children and youth. Professor Lipsey’s research has been supported by major federal funding agencies and foundations and recognized by awards from the university and major professional organizations. His published works include textbooks on program evaluation, meta-analysis, and statistical power as well as articles on applied methods and the effectiveness of school and community programs for youth. Professor Lipsey’s involvement in evaluation research began long ago in the doctoral psychology program at the Johns Hopkins University and includes graduate-level teaching at Claremont Graduate University and Vanderbilt, editorial roles with major journals in the field, directorship of several research centers dedicated to evaluation research, principal investigator on many evaluation research studies, consultation on a wide range of evaluation projects, and service on various national boards and committees related to applied social science.

Gary T. Henry

holds the Patricia and H. Rodes Hart Chair as a professor of public policy and education in the Department of Leadership, Policy and Organization at Peabody College, Vanderbilt University. He formerly held the Duncan MacRae '09 and Rebecca Kyle MacRae Distinguished Professorship of Public Policy in the Department of Public Policy and directed the Carolina Institute for Public Policy at the University of North Carolina at Chapel Hill. He has published extensively in top journals such as *Science*, *Educational Researcher*, *Journal of Policy Analysis and Management*, *Educational Evaluation and Policy Analysis*, *Journal of Teacher Education*, *Education Finance and Policy*, and *Evaluation Review*. Professor Henry's research has been funded by the Institute of Education Sciences, U.S. Department of Education, Spencer Foundation, Lumina Foundation, National Institute for Early Childhood Research, Walton Family Foundation, Laura and John Arnold Foundation, and various state legislatures, governor's offices, and agencies. Currently, he is leading the evaluation of the North Carolina school transformation initiative; the evaluation of Tennessee's school turnaround program; and the evaluation of the leadership pipeline in Hamilton County (Chattanooga), Tennessee. Dr. Henry serves as chair of the Education Systems and Broad Reform Research Scientific Review Panel for the Institute of Education Sciences, U.S. Department of Education. He has received the Outstanding Evaluation of the Year Award from the American Evaluation Association and the Joseph S. Wholey Distinguished Scholarship Award from the American Society for Public Administration and the Center for Accountability and Performance. In 2016, he was named an American Educational Research Association Fellow.

Chapter 1 What Is Program Evaluation and Why Is It Needed?

[What Is Program Evaluation?](#)

[Why Is Program Evaluation Needed?](#)

[Why Systematic Evaluation?](#)

[Systematic Program Evaluation](#)

[Application of Social Research Methods](#)

[The Effectiveness of Social Programs](#)

[Adapting to the Political and Organizational Context](#)

[Influencing Social Action to Improve Social Conditions](#)

[The Central Role of Evaluation Questions](#)

[The Purpose of the Evaluation](#)

[Program Improvement](#)

[Accountability](#)

[Knowledge Generation](#)

[Hidden Agendas](#)

[The Evaluator-Stakeholder Relationship](#)

[Criteria for Program Performance](#)

[The Five Domains of Evaluation Questions and Methods](#)

[Need for the Program: Needs Assessment](#)

[Assessment of Program Theory and Design](#)

[Assessment of Program Process](#)

[Effectiveness of the Program: Impact Evaluation](#)

[Cost Analysis and Efficiency Assessment](#)

[The Interplay Among the Evaluation Domains](#)

[Summary](#)

[Key Concepts](#)

Program evaluation is the systematic assessment of programs designed to improve social conditions and our individual and collective well-being. Programs are designed to address social problems, but most social problems resist efforts to remedy them. To answer key questions about the performance of such programs, evaluators apply social science research methods to provide answers to stakeholders. To be effective, a social program must correctly diagnose the problem it is intended to address, adopt a feasible design capable of ameliorating the problem, be well implemented in a manner consistent with the design, actually improve the outcomes for the population targeted by the program, and do so at an acceptable cost to

society. Different domains of program evaluation address questions related to each of these aspects of social programs using concepts and methods appropriate to those questions.

This book is rooted in the tradition of scientific study of social problems—a tradition that has aspired to improve the quality of social conditions and our physical environment and enhance our individual and collective well-being through the systematic creation and application of knowledge. Although the terms *program evaluation* and *evaluation research* are relatively recent inventions, the activities we will consider under these rubrics are not. They can be traced to the very beginnings of modern science. Three centuries ago, as Cronbach and colleagues (1980) point out, Thomas Hobbes and his contemporaries tried to use numerical measures to assess social conditions and identify the causes of mortality, morbidity, and social disorganization. Since the latter part of the 20th century, the resistance of many social problems to efforts to bring about change for the better and developments in empirical social sciences have combined to make program evaluation an important and commonplace undertaking.

What Is Program Evaluation?

Our focus is on social programs, also referred to as social interventions, especially human service programs in such areas as health, education, employment, housing, community development, poverty, criminal justice, and international development. At various times, policymakers, funding organizations, planners, program managers, taxpayers, or program clientele need to distinguish worthwhile social programs from ineffective ones, or perhaps launch new programs or revise existing ones so that the programs may achieve better outcomes. Informing and guiding the relevant stakeholders in their deliberations and decisions about such matters is the work of program evaluation. (Note that throughout this book we use the terms *evaluation*, *program evaluation*, and *evaluation research* interchangeably.)

Although this text emphasizes evaluation of social programs, evaluation research is not restricted to that arena. The broad scope of program evaluation can be seen in the evaluations of the U.S. Government Accountability Office (GAO), which have covered the procurement and testing of military hardware, quality control for drinking water, the maintenance of major highways, the use of hormones to stimulate growth in cattle, and other organized activities far afield from human services. Indeed, the techniques described in this text are useful in virtually all spheres of activity in which issues are raised about the effectiveness of organized social action. For example, the mass communication and advertising industries use essentially the same approaches in developing media programs and marketing products. Political candidates develop their campaigns by evaluating the voter appeal of different strategies. Consumer products are tested for performance, durability, and safety. This list of examples could be extended indefinitely.

To illustrate the evaluation of social programs more concretely, we offer below a few examples of diverse programs with different aims that have been evaluated in various settings and social sectors.

- In 2010, malaria was responsible for 1 million deaths per year worldwide according to the World Health Organization, and in Kenya it was responsible for one quarter of all children's deaths. Bed nets treated with insecticide have been shown to be effective in reducing maternal anemia and infant mortality, but in Kenya fewer than 5% of children and 3% of

pregnant women slept under them. In 16 Kenyan health clinics, pregnant women were randomly given an opportunity to obtain bed nets at no cost instead of the regular price. The acquisition and use of bed nets increased by 75% when they were free compared with the regular cost of 75 cents. In part because of the availability and use of bed nets, deaths attributable to malaria have been reduced by 29% since 2010 (Cohen & Dupas, 2010).

- Since the initiation of federal requirements for monitoring students' proficiency in reading, mathematics, and science as well as graduation rates, the issue of chronically low performing schools has garnered much public attention. In Tennessee some of the lowest performing schools were taken into a special district controlled by the state. Others were placed in special "district-within-districts," known as iZones, and granted greater autonomy and additional resources. In the first 3 years of operation, an evaluation showed that student achievement increased in the iZone schools, but not in the schools taken over by the state, which were run primarily by charter school organizations (Zimmer, Henry, & Kho, 2017).
- Acceptance and commitment theory (ACT) is a treatment program for individuals who engage in aggressive behavior with their domestic partners. Delivered in a group format, ACT targets such problematic characteristics of abusive partners as low tolerance for emotional distress, low empathy for the abused partner, and limited ability to recognize emotional states. An evaluation of ACT compared outcomes for ACT participants with comparable participants in a general support-and-discussion group that met for the same length of time. Outcomes measured 6 months later showed that ACT participants reported less physical and psychological aggression than participants in the discussion group (Zarling, Lawrence, & Marchman, 2015).
- The threat of infectious disease is high in office settings where employees work in close proximity, with implications for absenteeism, productivity, and health care insurance claims. A large company in the American Midwest attempted to reduce these adverse effects by placing hand sanitizer wipes in each office and liquid hand sanitizer dispensers in high-traffic common areas. This intervention was implemented in two of the three office buildings on the company's campus, with the third and largest building held back for comparison purposes. They found that during the 1st year there were 24% fewer health care claims for preventable infectious diseases among the employees in the treated buildings than in the prior year, and no change for the employees in the untreated building. Those employees also had fewer absences from work, and an employee

survey revealed increases in the perception of company concern for employee well-being (Arbogast et al., 2016).

These examples illustrate the diversity of social interventions that have been systematically evaluated and the globalization of evaluation research. However, all of them involve one particular evaluation activity: evaluating the effects of programs on relevant outcomes. As we will discuss later, evaluation may also focus on the need for a program; its design, operation, and service delivery; or its efficiency.

Why Is Program Evaluation Needed?

Most social programs are well intentioned and take what seem like quite reasonable approaches to improving the problematic situations they address. If that were sufficient to ensure their success, there would be little need for any systematic evaluation of their performance. Unfortunately, good intentions and intuitively plausible interventions do not necessarily lead to better outcomes. Indeed, they can sometimes backfire, with what seem to be promising programs having harmful effects that were not anticipated. For example, the popular Scared Straight program, which spawned a television series that lasted for nine seasons, involved taking juvenile delinquents to see prison conditions and interact with the adult inmates in order to deter crime. However, evaluations of the program found that it actually resulted in increased criminal activity among the participants (Petrosino, Turpin-Petrosino, Hollis-Peel, & Lavenberg, 2013). This example and countless others show that the problems social programs attack are rarely ones easily influenced by efforts to resolve them. They tend to be complex, dynamic, and rooted in entrenched behavior patterns and social conditions resistant to change.

Under these circumstances, there are many ways for intervention programs to come up short. They may be based on an action theory (more about this later) that is not well aligned with the nature or root causes of the problem, or one that assumes an unrealistic process for changing the conditions it addresses. Furthermore, any program with at least some potential to improve the pertinent outcomes must be well enough implemented to achieve that potential. A service that is not delivered or is poorly delivered relative to what is intended has little chance of accomplishing its goals. With an inherently effective intervention strategy that is adequately implemented and then actually has the intended beneficial effects, there can still be issues that keep the program from being a complete success. For example, the program may also have effects in addition to those intended that are not beneficial, that is, adverse side effects. And there is the issue of cost, whether to government and ultimately taxpayers or to private sponsors. A program may produce the intended benefits, but at such high cost that it is not viable or sustainable. Or there may be alternative program strategies that would be equally effective at lower cost.

In short, there are many ways for a program to fail to produce the intended benefits without unanticipated negative side effects, or to do so in a sustainable,

cost-effective way. Good intentions and a plausible program concept are not sufficient. If they were, we could be confident that most social programs are effective at delivering the expected benefits without conducting any evaluation of their theories of action, quality of implementation, positive and adverse effects, or benefit-cost relationships. Unfortunately, that is not the world we live in. When programs are evaluated, it is all too common for the results to reveal that they are not effective in producing the intended outcomes. If those outcomes are worth achieving, it is especially important under these circumstances to identify successful programs. But it is equally important to identify the unsuccessful ones so that they may be improved or replaced by better programs. Assessing the effectiveness of social programs and identifying the factors that drive or undermine their effectiveness are the tasks of program evaluation.

Why Systematic Evaluation?

The subtitle of this evaluation text is “A Systematic Approach.” There are many approaches that might be taken to evaluate a social program. We could, for example, simply ask individuals familiar with the program if they think it is a good program. Or, we could rely on the opinions of experts who review a program and render judgment, rather the way sommeliers rate wine. Or, we could assess the status of the recipients on the outcomes the program addresses to see how well they are doing and somehow judge whether that is satisfactory. Although any of these approaches would be informative, none are what we mean by *systematic*. The next section of this chapter will discuss this in more detail, but for now we focus on the challenges any evaluation approach must deal with if it is to produce valid, objective answers to critical questions about the nature and effects of a program. It is those challenges that motivate a systematic approach to evaluation.

One such challenge is the relativity of program effects. With rare exceptions, some program participants will show improvement on the outcomes the program targets, such as less depression, higher academic achievement, obtaining employment, fewer arrests, and the like, depending on the focus of the program. But that does not necessarily mean these gains were caused by participation in the program. Improvement for at least some individuals is quite likely to have occurred anyway in the natural course of events even without the help of the program. Crediting the program with all the improvement participants make will generally overstate the program effects. Indeed, there may be circumstances in which participation in the program results in less gain than recipients would have made otherwise, such as in the Scared Straight example. Thus program effects must be assessed relative to the outcomes expected without program participation, and those are usually difficult to determine.

It follows that program effects are often hard to discern. Take the example of a smoking cessation program. If every participant is a 20-year smoker who has tried unsuccessfully multiple times to quit before joining such a program, and none of them ever smoke again afterward, it is not a great leap to interpret this as largely a program effect. It seems reasonably predictable that all of the participants would not have quit smoking in the absence of the program. But what if 60% start smoking again? Relapse rates are high for addictive

behaviors, but could there be a program effect in that high rate? Maybe 70% would start smoking again without the program. Or maybe only 50%. Most program effects are not black or white, but in the gray area where the influence of the program is not obvious.

A direct approach to this ambiguity would be to ask the participants if the program helped them. They will almost certainly have opinions to offer, but they will not be reliable informants about program effects. Those who have done well will likely give exaggerated credit to the program, but it is as much a matter of speculation for them as it is for evaluators to rule out the possibility that they would have done as well without the program. The clearest indication of this inclination for participants to credit the program for their successes is the ready availability of testimonials for virtually every program. Even programs found to be ineffective in rigorous evaluations can generally find participants who did well and will attribute their success to the program. It is simply very difficult for people to accurately account retrospectively for the factors that actually caused their behavior to change.

Alternatively, we might ask the program providers about how effective the program is. The line staff who deliver the services and interact directly with recipients certainly seem to be in a position to provide a good assessment of how well the program is working. Here, however, we encounter the problem of **confirmation bias**: the tendency to see things in ways favoring preexisting beliefs. Consider the medical practitioners in bygone eras who were convinced by the evidence of their own eyes and the wisdom of their clinical judgment that treatments we now know to be harmful, such as bloodletting and mercury therapy, were actually effective. They did not intend to harm their patients, but they believed in those treatments and gave much greater weight in their assessment to patients who recovered than those who did not. Similarly, program providers generally believe the services they provide are beneficial, and confirmation bias nudges them to high awareness of evidence consistent with that belief and to discount contrary evidence.

The approaches to evaluating the performance of a program that may seem most natural and straightforward, therefore, cannot be counted on to provide a valid assessment. If program evaluation is to reach valid conclusions about program performance, systematic methods structured to avoid bias and misrepresentation as much as possible must be used.

Systematic Program Evaluation

We begin with the definition of program evaluation that guides the orientation of this text and then elaborate on each component of this definition to highlight the major themes we believe are integral to the practice of program evaluation.

Program evaluation is the application of social research methods to systematically investigate the effectiveness of social intervention programs in ways that are adapted to their political and organizational environments and are designed to inform social action to improve social conditions.

One of the pioneers of systematic program evaluation, who developed and refined many of the practices and methods used in the field today, was the first author of this text, Peter H. Rossi. Rossi, who passed away in 2006, was a leading sociologist who served on the faculty of Harvard, the University of Chicago, Johns Hopkins, and the University of Massachusetts–Amherst and conducted research on social problems and evaluated social programs. His vision for systematic program evaluation and some of his contributions to the field are noted in [Exhibit 1-A](#).

Application of Social Research Methods

The concept of evaluation entails, on one hand, a description of the performance of the entity being evaluated and, on the other, some standards or criteria for judging that performance (see [Exhibit 1-B](#)). It follows that a central task of the program evaluator is to construct a valid description of program performance in a form that permits comparison with applicable criteria. Failing to describe program performance with a reasonable degree of validity may distort a program's accomplishments, deny it credit for its successes, or overlook shortcomings for which it should be accountable. Moreover, an acceptable description of program performance must be detailed and precise. An unduly vague or equivocal description will make it difficult to determine with confidence whether the performance actually meets the appropriate standard.

Exhibit 1-A Peter H. Rossi: An Evaluation Champion and Legendary Evaluator



The major reason why public social programs fail is that effective programs are difficult to design. . . . The major sources of program design failures are: (a) incorrect understanding of the social problem being addressed, (b) interventions that are inappropriate, and (c) faulty implementation of the intervention.

. . . I believe that we can make the following generalization: **The findings of the majority of evaluations purporting to be impact assessments are not credible.**

They are not credible because they are built upon research designs that cannot be safely used for impact assessments. I believe that in most instances, the fatal design defects are not possible to remedy within the time and budget constraints faced by the evaluator.

Source: Rossi (2003).

One example of Peter Rossi's systematic approach to evaluation was his application of sampling theory and social science data collection methods to assess the needs of the homeless in Chicago. He became the first to obtain a credible estimate of the number of homeless individuals in the city, distinguishing residents of shelters and those living on the streets. For counts of shelter residents, his research team visited all the homeless shelters in Chicago for 2 weeks in the fall and 2 weeks in the winter. To collect additional data, he sampled shelters and

residents within them for participation in a survey. For the homeless living on the streets, he sampled city blocks and then canvased the homeless individuals on each sampled block between 1 a.m. and 6 a.m. to reduce duplicate counts of shelter residents. The researchers were accompanied by out-of-uniform police officers for their safety, and respondents were paid for their participation in the study. Rossi's research revealed that the homeless population was much smaller than claimed by advocates for the homeless and that it had changed to include more women and minorities than in earlier homeless populations. He found that structural factors, such as the decline of jobs for low-skilled individuals, contributed to homelessness, but it was personal factors like alcoholism and physical health problems that separated the homeless from other extremely poor individuals. This is but one example of his influential contributions to evaluation, which also included evaluations of federal food programs, public welfare programs, and anticrime programs.

Source: Rossi (1990).

Exhibit 1-B The Two Arms of Evaluation

Evaluation is the process of determining the merit, worth, and value of things, and evaluations are the products of that process. . . . Evaluation is not the mere accumulation and summarizing of data that are clearly relevant for decision making, although there are still evaluation theorists who take that to be its definition. . . . In all contexts, gathering and analyzing the data that are needed for decision making—difficult though that often is—comprises only one of the two key components in evaluation; absent the other component, and absent a procedure for combining them, we simply lack anything that qualifies as an evaluation. *Consumer Reports* does not just test products and report the test scores; it (i) *rates or ranks* by (ii) *merit or cost-effectiveness*. To get to that kind of conclusion requires an input of something besides data, in the usual sense of that term. The second element is required to get to conclusions about merit or net benefits, and it consists of evaluative premises or standards. . . . A more straightforward approach is just to say that evaluation has two arms, only one of which is engaged in data-gathering. The other arm collects, clarifies, and verifies relevant values and standards.

Source: Scriven (1991, pp. 1, 4–5).

Social research methods and the accompanying standards of methodological quality have been developed and refined explicitly for the purpose of constructing sound factual descriptions of social phenomena. In particular, contemporary social science techniques of systematic observation, measurement, sampling, research design, and data analysis represent highly refined procedures for producing valid, reliable, and precise characterizations of social behavior. Social research methods thus provide an especially appropriate approach to the task of describing program performance in ways that will be as credible and defensible as possible.

Regardless of the type of social intervention under study, therefore, evaluators will typically use social research procedures for gathering, analyzing, and interpreting evidence about the performance of a program. This is not to say, however, that we believe that program evaluation must use some particular

social research methods or combination of methods, whether quantitative or qualitative, experimental or ethnographic, positivist or naturalist. Nor does this commitment to the methods of social science mean that we think current methods are beyond improvement. Evaluators must often innovate and improvise as they attempt to find ways to gather credible, compelling evidence about social programs. In fact, evaluators have made many novel contributions to methodological development in applied social research in their quest to improve the evidence they can provide about social programs and their effectiveness.

Nor does this view imply that methodological quality is necessarily the most important aspect of an evaluation or that only the highest technical standards, without compromise, are always appropriate. As Carol Weiss (1972) observed long ago, social programs are inherently inhospitable environments for research purposes. The people operating social programs tend to focus attention on providing the services they are expected to provide to the members of the target population specified to receive them. Gathering data is often viewed as a distraction from that central task. The circumstances surrounding specific programs and the issues the evaluator is called on to address frequently compel them to adapt textbook methodological standards, develop innovative methods, and make compromises that allow for the realities of program operations and the time and resources allocated for the evaluation. The challenges to the evaluator are to match the research procedures to the evaluation questions and circumstances as well as possible and, whatever procedures are used, to apply them at the highest standard possible to those questions and circumstances.

The Effectiveness of Social Programs

Social programs are generally undertaken to “do good,” that is, to ameliorate social problems or improve social conditions. It follows that it is appropriate for the parties who invest in social programs to hold them accountable for their contribution to the social good. Correspondingly, any evaluation of such programs worthy of the name must evaluate—that is, judge—the quality of a program’s performance as it relates to some aspect of its effectiveness in producing social benefits. More specifically, the evaluation of a program generally involves assessing one or more of five domains: (a) the need for the program, (b) its design and theory, (c) its implementation and service delivery, (d) its outcome and impact, and (e) its efficiency (more about these domains later in the chapter).

Adapting to the Political and Organizational Context

Program evaluation is not a cut-and-dried activity like putting up a prefabricated house or checking a student's paper with a computer program that detects plagiarism. Rather, evaluators must tailor the evaluation to the particular program and its circumstances. The specific form and scope of an evaluation depend primarily on its purposes and audience, the nature of the program being evaluated, and, not least, the political and organizational context within which the evaluation is conducted. Here we focus on the last of these factors, the context of the evaluation.

The evaluation plan is generally organized around questions posed about the program by the **evaluation sponsor**, who commissions the evaluation, and other pertinent **stakeholders**: individuals, groups, or organizations with a significant interest in how well a program is working. These questions may be stipulated in specific, fixed terms that allow little flexibility, as in a detailed contract for evaluation services. However, it is not unusual for the initial questions to be vague, overly general, or phrased in program jargon that must be translated for more general consumption. Occasionally, the evaluation questions put forward are essentially pro forma (e.g., is the program effective?) and have not emerged from careful reflection regarding the relevant issues. In such cases, the evaluator must probe thoroughly to determine what the questions mean to the evaluation sponsor and stakeholders.

Equally important are the reasons the questions are being asked, especially the uses that are intended for the answers. An evaluation must provide information that addresses issues that matter for the key stakeholders and communicate it in a form that is usable for their purposes. For example, an evaluation might be designed one way if it is to provide information about the quality of service as feedback to the program director, who will use the results to incrementally improve the program, and quite another way if it is to provide information to a program sponsor, who will use it to decide whether to renew the program's funding.

These assertions assume that an evaluation would not be undertaken unless there was an audience interested in receiving and at least potentially using the findings. Unfortunately, evaluations are sometimes commissioned with little intention of using the findings. For instance, an evaluation may be conducted

solely because it is mandated by program funders and then used only to demonstrate compliance with that requirement. Responsible evaluators try to avoid being drawn into such situations of ritualistic evaluation. An early step in planning an evaluation, therefore, is an inquiry into the motivation of the evaluation sponsors, the intended purposes of the evaluation, and the uses to be made of the findings.

As a practical matter, an evaluation must also be tailored to the organizational makeup of the program. In designing an evaluation, the evaluator must take into account such organizational factors as the availability of administrative cooperation and support; the ways in which program files and data are kept and the access permitted to them; the character of the services provided; and the nature, frequency, duration, and location of the contact between the program and its clients. Once the evaluation is under way, modifications may be necessary in the types, quantity, or quality of the data collected as a result of unanticipated practical or political obstacles, changes in the operation of the program, or shifts in the interests of the stakeholders.

Influencing Social Action to Improve Social Conditions

We have emphasized that the role of evaluation is to provide answers to questions about a program that will be useful and will be used. This point is fundamental to evaluation: its purpose is to influence action. An evaluation, therefore, primarily addresses the audiences with the potential to make decisions and take action on the basis of the evaluation results. The evaluation findings may assist in making go/no-go decisions about specific program modifications or, perhaps, about initiation or continuation of entire programs. The evaluation may have direct effects on judgments of a program's value as part of an oversight process that holds the program accountable for results. Or it may have indirect effects in shaping the way program issues are framed and the nature of the debate about them.

Program evaluations may also have social action purposes beyond those of the particular programs being evaluated. What is learned from an evaluation of one program, say, a drug use prevention program at a particular high school, says something about the whole category of similar programs. Many of the parties involved with social interventions must make decisions and take action that relates to types of programs rather than individual programs. A congressional committee may debate the merits of privatizing public education, a state correctional department may consider instituting community-based substance abuse treatment programs, or a philanthropic foundation may deliberate about whether to provide contingent incentives to parents that encourage their children to remain in school. The body of evaluation findings for programs of each of these types is very pertinent to discussions and decisions at this broader level.

One important form of evaluation research is conducted on **demonstration programs**, which are social intervention projects designed and implemented explicitly to test the value of an innovative program concept. In such cases, the findings are significant because of what they reveal about the program concept and how promising it is for broader implementation. Another significant evaluation-related activity is the integration of the findings of multiple evaluations of a particular type of program into a synthesis that can inform policy making and program planning. Whether focused on an individual

program or a collection of programs, the common denominator in all evaluation research is that it is intended to be both useful and used, either directly and immediately or as an incremental contribution to a cumulative body of practical knowledge.

The Central Role of Evaluation Questions

One of the most challenging aspects of evaluation is that there is no one-size-fits-all approach. Every evaluation situation has a different and unique profile of characteristics. A good evaluation design is one that adapts the evaluator's repertoire of approaches, techniques, and concepts to the program circumstances in a way that yields credible and useful answers to the questions that motivate it. The nature of those evaluation questions and the way they are developed and formulated are not only the starting point for any program evaluation but the organizing themes around which the evaluation is structured. In this section we review some of the key features of evaluation questions and the factors that shape them.

The Purpose of the Evaluation

Evaluations are initiated for many reasons. They may be intended to help management improve a program; support advocacy by proponents or critics; gain knowledge about the program's effects; provide input to decisions about the program's funding, structure, or administration; or respond to political pressures. One of the first determinations the evaluator must make to identify the most relevant evaluation questions is the purpose of the evaluation. This is not always a simple matter. A statement of the purposes may accompany the request for an evaluation, but those announced purposes rarely tell the whole story and sometimes are only rhetorical. The evaluator often must dig deeper to determine who wants the evaluation, what they want, and why they want it. There is no cut-and-dried method for doing this, but it is usually best to approach the task the way a journalist would dig out a story. The evaluator can examine source documents, interview key informants with different vantage points, and uncover pertinent history and background. Generally, the purposes of the evaluation will relate mainly to program improvement, accountability, or knowledge generation, but sometimes quite different motivations are in play.

Program Improvement

An evaluation intended to furnish information for guiding program improvement is called a **formative evaluation** (Scriven, 1991) because its purpose is to help form or shape the program to perform better. The audiences for formative evaluations typically are program planners, administrators, oversight boards, or funders with an interest in optimizing the program's effectiveness. The information desired may relate to the need for the program, the program's design, its implementation, its impact, or its costs, but often tends to focus on program operations, service delivery, and take-up of services by the program's target population. The evaluator in this situation will usually work closely with program management and other stakeholders in designing, conducting, and reporting the evaluation. Evaluation for program improvement characteristically emphasizes findings that are timely, concrete, and immediately useful. Correspondingly, the communication between the evaluator and the respective audiences may occur regularly throughout the evaluation and can be relatively informal.

Accountability

The investment of social resources such as taxpayer dollars by human service programs is justified by the presumption that the programs will make beneficial contributions to society. Program managers are thus expected to use resources effectively and efficiently and actually produce the intended benefits. An evaluation conducted to determine whether these expectations are met is called a **summative evaluation** (Scriven, 1991) because its purpose is to render a summary judgment on the program's performance. The findings of summative evaluations are usually intended for decision makers with major roles in program oversight, for example, the funding agency, governing board, legislative committee, political decision makers, or organizational leaders. Such evaluations may influence significant decisions about the continuation of the program, allocation of resources, restructuring, or legislative action. For this reason, they require information that is sufficiently credible under scientific standards to provide a confident basis for action and to withstand criticism aimed at discrediting the results. The evaluator may be expected to function relatively independently in planning, conducting, and reporting the evaluation, with stakeholders providing input but not participating directly in decision making. In these situations, it may be important to avoid premature or careless conclusions, so communication of the evaluation findings may be relatively formal, rely chiefly on written reports, and occur primarily at the end of the evaluation.

Knowledge Generation

Some evaluations are undertaken to describe the nature and effects of an intervention as a contribution to knowledge. For instance, an academic researcher might initiate an evaluation to test whether a program designed on the basis of theory, say, a behavioral nudge to undertake a socially desirable behavior, is workable and effective. Similarly, a government agency or private foundation may mount and evaluate a demonstration program to investigate a new approach to a social problem, which, if successful, could then be implemented more widely. Because evaluations of this sort are intended to make contributions to the social science knowledge base or be a basis for significant program innovation, they are usually conducted using the most rigorous methods feasible. The audience for the findings will include the sponsors of the research as well as a broader audience of interested scholars and

policymakers. In these situations, the findings of the evaluation are most likely to be disseminated through scholarly journals, research monographs, conference papers, and other professional outlets.

Hidden Agendas

Sometimes the true purpose of the evaluation, at least for those who initiate it, has little to do with actually obtaining information about the program's performance. Program administrators or boards may launch an evaluation because they believe it will be good for public relations and might impress funders or political decision makers. Occasionally, an evaluation is commissioned to provide a rationale for a decision that has already been made behind the scenes to terminate a program, fire an administrator, or the like. Or the evaluation may be commissioned as a delaying tactic to appease critics and defer difficult decisions.

Virtually all evaluations involve some political maneuvering and public relations, but when these are the principal purposes, the prospective evaluator is presented with a difficult dilemma. The evaluation must either be guided by the political or public relations purposes, which will likely compromise its integrity, or focus on program performance issues that are of little real interest to those commissioning the evaluation and may even be threatening. In either case, the evaluator is well advised to try to avoid such situations.

The Evaluator-Stakeholder Relationship

Every program is necessarily a social structure in which various individuals and groups engage in the roles and activities that constitute the program. In addition, every program is a nexus in a set of political and social relationships among those with involvement or interest in the program, such as relevant decision makers, competing programs, and advocacy groups. The nature of the evaluator's relationship with these and other stakeholders who may participate in the evaluation or have an interest in it will shape the way the evaluation questions are framed. The primary stakeholders potentially influential in this process may include the following:

Decision makers: Persons responsible for deciding whether the program is to be initiated, continued, discontinued, expanded, modified, restructured, or curtailed.

Program sponsors: Individuals with positions of responsibility in public agencies or private organizations that initiate and fund the program; they may overlap with decision makers.

Evaluation sponsors: Individuals in public agencies or private organizations who initiate and fund the evaluation (the evaluation sponsors and program sponsors may be the same).

Target participants: Persons, households, or other units that are intended to receive the intervention or services being evaluated.

Program managers: Personnel responsible for overseeing and administering the intervention program.

Program staff: Personnel responsible for delivering the program services or functioning in supporting roles.

Program competitors: Organizations or groups that compete with the program. For instance, a private organization receiving public funds to operate charter schools will be in competition with public schools also supported by public funds.

Contextual stakeholders: Organizations, groups, and individuals in the environment of a program with interests in what the program is doing or what happens to it (e.g., other agencies or programs, journalists, public officials, advocacy organizations, citizens' groups in the jurisdiction in which the program operates).

Evaluation and research community: Evaluation professionals who read evaluations and review their technical quality and credibility along with

researchers who work in areas related to that type of program.

The most influential stakeholder will typically be the evaluation sponsor, the agent that initiates the evaluation, usually provides the funding, and makes decisions about how and when it will be done and who will do it. Various relationships with the evaluation sponsor and other stakeholders are possible and will depend largely on the sponsor's preferences and whatever negotiation takes place with the evaluator. The evaluator's relationship to stakeholders is so influential for shaping the evaluation process that a special vocabulary has arisen to describe the major variants.

In an **independent evaluation**, the evaluator has the primary responsibility for developing the evaluation questions in collaboration with key stakeholders, conducting the evaluation, and disseminating the results. The evaluator may initiate and direct the evaluation quite autonomously, as when a social scientist undertakes an evaluation for purposes of knowledge generation with research funding that leaves the particulars to the researcher's discretion. More often, the independent evaluator is commissioned by a sponsoring agency that stipulates the purposes and nature of the evaluation but leaves it to the evaluator to do the detailed planning and conduct the evaluation. For instance, program funders often commission evaluations by publishing a request for proposals or applications, to which evaluators respond with statements of their capability, proposed design, budget, and time line, as requested. The evaluation sponsor then selects an evaluator from among those responding and establishes a contractual arrangement for the agreed-on work. In such cases, however, the evaluator nonetheless generally confers with a range of stakeholders to give them some influence in shaping the evaluation.

A **participatory or collaborative evaluation** is organized as a team project with the evaluator and representatives of one or more stakeholder groups jointly making decisions about the evaluation and how it is conducted. The participating stakeholders are directly involved in formulating the evaluation questions, and planning, conducting, and analyzing the data collected for the evaluation in collaboration with the evaluator. The evaluator's role might range from project leader or coordinator to that of resource person called on only as needed. Variations on this form of relationship are typical for internal evaluators who are part of the organization whose program is being evaluated. In such cases, the evaluator generally works closely with management in formulating the evaluation questions and planning and conducting the evaluation. One well-

known form of participatory evaluation is Patton's (2008) utilization-focused evaluation. Patton's approach emphasizes close collaboration with the individuals who will use the evaluation findings to ensure that it is responsive to their needs and produces information they can and will actually use.

In an **empowerment evaluation**, the evaluator-stakeholder relationship is participatory and collaborative. In addition, however, the evaluator's role includes consultation and facilitation directed toward democratic participation and building the capacities of the participating stakeholders to conduct evaluations on their own, to use the results effectively for advocacy and change, and to take ownership of a program that affects their lives. For instance, some recipients of program services may be asked to take a primary role in planning, setting priorities, collecting information, and interpreting the results of the evaluation. The evaluation process in this arrangement, therefore, is directed not only at producing informative and useful findings but also at enhancing the development and political influence of the participants. As these themes imply, empowerment evaluation most appropriately includes stakeholders who otherwise have little power in the context of the program, usually the program recipients or intended beneficiaries. In their most recent contribution, three pioneers of empowerment evaluation document examples in contexts as diverse as a tobacco prevention program and an organizational transformation initiative that have used this approach (Fetterman, Kaftarian, & Wandersman, 2015).

Criteria for Program Performance

Beginning a study with a set of research questions is customary in the social sciences (often framed as hypotheses). What distinguishes evaluation questions is that they have to do with performance and are associated, at least implicitly, with some criteria by which that performance can be judged. When program managers or evaluation sponsors ask such things as “Are we targeting the right client population?” or “Do our services benefit the recipients?” they are not only asking for a description of the program’s performance, they are also asking if that performance is good enough according to some standard or judgment.

One implication of this distinctive feature of evaluation is that good evaluation questions will, when possible, convey the applicable **performance criterion** or standard as well as the performance dimension that is at issue. Thus, evaluation questions may be much like this: “Does the program serve at least 75% of the individuals eligible to receive the services?” (by some explicit eligibility criteria) or “Do the majority of those who receive the employment services get jobs within 30 days of the conclusion of training that they keep at least 3 months?” To be meaningful, there should be some rationale for the standard that is related to the ability of the program to accomplish its overall goal of improving the target social conditions.

The applicable performance criteria may take different forms for various dimensions of program performance ([Exhibit 1-C](#)). In some instances, there are established professional standards that are applicable to program performance. This is particularly likely in medical and health programs, in which practice guidelines and managed care standards may be relevant. Perhaps the most common criteria are those based directly on program design, goals, and objectives. In this case, program officials and sponsors identify certain desirable accomplishments as the program aims. Often these statements are not very specific with regard to the nature or level of program performance they represent. One of the goals of a shelter for battered women, for instance, might be to “empower women to take control of their own lives.” Although reflecting commendable values, this statement gives no indication of the tangible manifestations of such empowerment that would constitute attainment of this goal. Considerable discussion with stakeholders may be necessary to translate such statements into mutually acceptable terminology that describes the intended outcomes concretely, identifies the observable indicators of those

outcomes, and specifies the level of accomplishment that would be considered a success in accomplishing the stated goal.

Some program objectives, on the other hand, may be very specific. These often come in the form of administrative objectives adopted as targets according to past experience, benchmarking against the experience of comparable programs, a judgment of what is reasonable and desirable, or maybe only an informed guess as to what is needed. Examples of administrative objectives may be to complete intake for 90% of the referrals within 30 days, to have 75% of the clients complete the full term of service, to have 85% “good” or “outstanding” ratings on a client satisfaction questionnaire, to provide at least three appropriate services to each person under case management, and the like. There is typically some arbitrariness in these criterion levels. But if they are administratively stipulated, can be established through stakeholder consensus, represent attainable targets for improvement over past practice, or can be supported by evidence of levels associated with positive outcomes, they may be quite serviceable in the formulation of evaluation questions and interpretation of the subsequent findings. However, it is not generally wise for the evaluator to press for specific statements of target performance levels if the program does not have them or cannot readily and confidently develop them.

Establishing a performance criterion can be particularly difficult when the performance dimension in an evaluation question involves outcome or impact issues. Program stakeholders and evaluators alike may have little idea about how much change on an outcome (e.g., frequency of alcohol or drug use) is large enough to have practical significance. In practice, the standard for performance is often set in relation to the outcome expected in the absence of the program and a related judgment about whether the program has improved on that at all. By default, these judgments are often made on the basis of statistical criteria, that is, whether the measured effects are statistically significant. This is a poor practice for reasons that will be more fully examined in [Chapter 9](#). Statistical criteria have no intrinsic relationship to the practical significance of a change on an important outcome and can be misleading. A juvenile delinquency program that is found to have the statistically significant effect of lowering subsequent reoffense rates by 2%, for example, may not make a large enough difference to be judged worthwhile relative to its costs.

The standards by which program performance may be judged in an evaluation include the following:

The needs or wants of the target population

Stated program goals and objectives

Professional standards

Customary practice; norms for other programs

Legal requirements

Ethical or moral values; social justice, equity

Past performance; historical data

Targets set by program managers

Expert opinion

Preintervention baseline levels for the target population

Conditions expected in the absence of the program (counterfactual)

Cost or relative cost

The Five Domains of Evaluation Questions and Methods

A carefully developed set of **evaluation questions** gives structure to an evaluation, leads to appropriate and thoughtful planning, and serves as a basis for discussions about who is interested in the answers and how they are to be used. Although appropriate evaluation questions will be rather specific to the program to be evaluated, it is useful to recognize that they generally fall into categories according to the program issues they address. Five such domains of evaluation questions can be distinguished:

Need for the program: Questions about the social conditions a program is intended to ameliorate and the need for the program.

Program theory and design: Questions about program conceptualization and design.

Program process: Questions about program operations, implementation, service delivery, and the way recipients experience the program services.

Program impact: Questions about program change in the targeted outcomes and the program's impact on those changes.

Program efficiency: Questions about program cost and cost-effectiveness.

Evaluators have developed concepts and methods for addressing the kinds of questions in each of these categories, and those combinations of questions, concepts, and methods constitute the primary domains of evaluation practice. Below we provide an overview of each of those five domains.

Need for the Program: Needs Assessment

The primary rationale for a social program is to alleviate a social problem. The impetus for a new program to increase adult literacy, for example, is likely to be recognition that a significant proportion of persons in a given population are deficient in reading skills. Similarly, an ongoing program may be justified by the persistence of a social problem: Driver education in high schools receives public support because of the continuing high rates of automobile accidents among adolescent drivers.

One important form of evaluation, therefore, assesses the nature, magnitude, and distribution of a social problem; the extent to which there is a need for intervention; and the implications of these circumstances for the design of the intervention. These diagnostic activities are referred to as **needs assessment** in the evaluation field (Altschuld & Kumar, 2010; Watkins, Meiers, & Visser, 2012) but overlap with what is called social epidemiology and social indicators research in other fields. Critical to the process of conducting a needs assessment is determination of the gap between the current social condition and the condition judged to be acceptable to society or a particular community.

Examples of the kinds of questions addressed by needs assessment, stated in summary form, are as follows:

- What are the nature and magnitude of the problem to be addressed?
- What are the characteristics of the population in need?
- What are the needs of the population? What has created that need?
- What kinds of assistance might address those needs? What outcomes would be desirable?
- What characteristics of the population in need would influence the ability to provide assistance or the way in which it should be provided?

Needs assessment to provide information about the nature of the social condition at issue and the implications for the ways in which it might be effectively addressed is often a first step in planning a new program. Needs assessment may also be appropriate to examine whether an established program is responsive to the current needs of its target population and provide guidance for improvement. [Exhibit 1-D](#) provides an example of one of the several

approaches that can be taken. [Chapter 2](#) discusses the various aspects of needs assessment in detail.

Exhibit 1-D Assessing the Needs of Older Caregivers for Young Persons Infected or Affected by HIV or AIDS

In South Africa, many aspects of the reduction of the incidence of HIV infection and AIDS and management of care for HIV-infected individuals and those with AIDS have been the focus of government interventions. However, the needs of older persons who are the primary caregivers for children or grandchildren affected by HIV or AIDS had not been previously assessed. In one arm of a mixed-methods study, evaluators selected and surveyed individuals, 50 years of age or older who were giving care to younger persons who received HIV- or AIDS-related services from one of seven randomly selected nongovernmental organizations (NGOs) in three of South Africa's nine provinces. In addition to the survey data, the evaluators selected 10 survey respondents for in-depth interviews and 9 key informants who managed government HIV/AIDS interventions or NGO programs.

Quantitative data were collected to assess the extent of the problem of caregiving by older persons, and qualitative data were collected to understand the burden of caregiving on the caregivers and to identify areas of need for formal support. A semistructured survey instrument was tested, refined, piloted, and then used to assess demographic and household data, health status, knowledge and awareness of HIV and AIDS, caregiving to persons living with the disease, caregiving to children and orphaned grandchildren, and support received from the government and other community institutions. Interview schedules were used to interview a purposive sample of caregivers, government officials, and managers of NGOs.

The evaluators collected data on the challenges and support needs of older caregivers and the gaps in public policy responses to the burden of care on those caregivers. The 305 respondents were 91% older women with a mean age of 66 years. Results highlighted that caregiving was largely feminized, and a majority of the caregivers (59%) relied on informal support from NGOs and family members. Lack of formal support was identified across all three provinces. The study was used to formulate a policy framework to inform the design and implementation of policy and programmatic responses aimed at supporting the caregivers.

Source: Adapted from Petros (2011).

Assessment of Program Theory and Design

Given a recognized problem and need for intervention, another domain for evaluation involves questions about the design of the program or intervention that is expected to address that need. The conceptualization and operational plan of a program must reflect valid assumptions about the nature of the problem and represent a feasible approach to reducing the gap between current and acceptable levels of the problematic condition. This program plan may not be written out in detail, but exists nonetheless as a shared conceptualization among the principal stakeholders. The critical part of program design consists of assumptions and expectations about how the program should operate in order to have the intended effects and is referred to as the program theory or theory of action. If this theory is faulty, the intervention will fail no matter how elegantly it is conceived or how well it is implemented.

Examples of questions that may guide an **assessment of program theory and design** in summary form are the following:

- What outcomes does the program intend to affect, and how do they relate to the nature of the problem or conditions the program aims to change?
- What is the theory of action that supports the expectation that the program can have the intended effects on the targeted outcomes?
- Is the program directed to an appropriate population, and does it incorporate procedures capable of recruiting and sustaining their participation in the program?
- What services does the program intend to provide, and is there a plausible rationale for the expectation that they will be effective?
- What delivery systems for the services are to be used, and are they aligned with the nature and circumstances of the target population?
- How will the program be resourced, organized, and staffed, and does that scheme provide an adequate platform for recruiting and serving the target population?

This type of assessment involves, first, describing the program theory in explicit and detailed form, often in the form of a logic model or a theory of behavioral or social change rooted in social science. Logic models are generally organized around the inputs required for a program, the actions or activities to be undertaken, the outputs from those activities, and the immediate, intermediate,

and ultimate outcomes the program aims to influence (Knowlton & Phillips, 2013). Programs designed around social science concepts are often drawn from theories of behavioral change, such as outsider theory that begins with dissatisfaction with one's current state and continues through anticipation of the benefits of changing behavior to the adoption of new behavior (Pawson, 2013). Once the program theory is formulated, various approaches are used to examine how reasonable, feasible, ethical, and otherwise appropriate it is. The sponsors of this form of evaluation are generally funding agencies or other decision makers attempting to launch a new program. [Exhibit 1-E](#) provides an example and [Chapter 3](#) offers further discussion of program theory and design as well as the ways in which it can be evaluated.

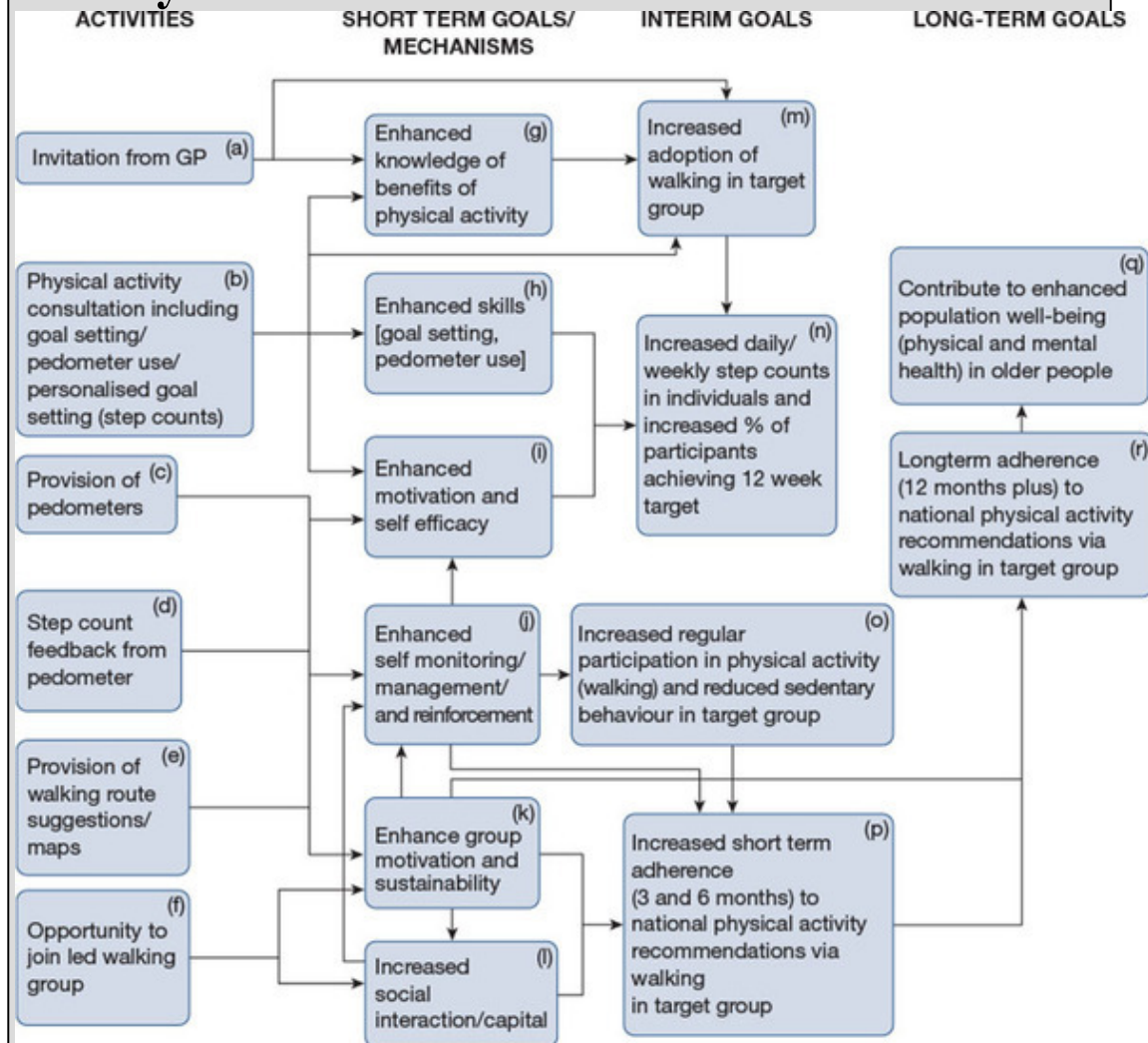
Assessment of Program Process

Given a plausible theory about how to intervene to ameliorate an accurately diagnosed social problem, a program must still be implemented well to have a reasonable chance of actually improving the situation. It is not unusual to find that programs are not implemented and executed according to their intended designs. A program may be poorly managed, compromised by political interference, or designed in ways that are impossible to carry out. Sometimes appropriate personnel are not available, facilities or resources are inadequate, or program staff lack motivation, expertise, or training. Possibly the intended program participants do not exist in the numbers required, cannot be identified precisely, or are difficult to engage.

Exhibit 1-E Assessing the Program Theory for a Physical Activity Intervention

Research indicates that physical activity can improve mental well-being, help with weight maintenance, and reduce the risk for chronic diseases such as diabetes. Despite such evidence, it was reported in 2011 that 67% of women and 55% of men in Scotland did not reach the minimum level of activity needed to attain such health benefits. As a result, an intervention known as West End Walkers 65+ (WEW65+) was developed in Scotland to increase walking and reduce sedentary behavior in adults older than 65 years. The design of the intervention relied heavily on empirically supported theories underlying behavioral change and prior activity interventions that had demonstrated effectiveness. Before implementation, the intervention design and underlying theory, depicted below, was assessed as part of a pilot and feasibility assessment of the program.

Theory for WEW65+ intervention



While assessing the program theory, the evaluators examined the underlying assumptions and the triggers for the psychological mechanisms expected lead to achieving the outcomes goals set for the intervention. They confirmed the reasonableness of assumptions such as the focus on an older population of adults, the appropriateness of walking as a sufficient physical activity to enhance health outcomes and reduce sedentariness, and the likelihood that information provided in a clinical setting to influence attitudes and behaviors. They also noted the addition of a program activity based on previously tested behavioral theory—a physical activity consultation to enhance the participants knowledge of the benefits of walking and enhance their motivation and self-efficacy—to the intervention design.

Source: Adapted from Blamey, Macmillan, Fitzsimons, Shaw, and Mutrie (2013).

A basic and widely used form of evaluation, **assessment of program process**, evaluates the fidelity and quality of a program's implementation. Such process assessments may be done as a freestanding evaluation of the activities and

operations of the program, commonly referred to as a process evaluation or an implementation assessment. When the process evaluation is an ongoing function that occurs regularly, it will usually be referred to as **program monitoring**. A program monitoring function may also include information about the status of program participants on targeted outcomes after they have completed the program and thus also include **outcome monitoring**. Process evaluation investigates how well the program is operating. It might examine how consistent the services actually delivered are with the design for the program, whether services are delivered to appropriate recipients, how well service delivery is organized, the effectiveness of program management, the use of program resources, the well-being of participants after receipt of program services, and other such matters ([Exhibit 1-F](#) provides an example). Examples of the kind of evaluation questions that guide process evaluations are:

- Are the intended services being delivered to the intended persons?
- Are administrative and service objectives being met?
- Are there eligible but unserved persons the program is not reaching?
- Once beginning service, do sufficient numbers of participants complete service?
- Are the participants satisfied with the services?
- Are the participants doing well in the ways intended after receipt of the program services?
- Are administrative, organizational, and personnel functions managed well?

Exhibit 1-F Assessing the Implementation Fidelity and Process Quality of a Youth Violence Prevention Program

After a pilot study proved successful, a community-level violence prevention and positive youth development program, Youth Empowerment Solutions (YES), was rolled out, and a process evaluation was conducted to measure implementation fidelity and quality of delivery. The process evaluation was conducted in 12 middle and elementary schools in Flint, Michigan, and surrounding Genesee County. Data were collected from 25 YES groups from 12 schools over 4 years. Four groups were eliminated from the analysis because of incomplete data. Data collection covered the measurement of implementation fidelity, the dose delivered to participants, the dose received from participants, and program quality. The evaluators summarized multiple methods adopted to measure each component in the table below.

	Participant- Teacher Interaction: Observation	Core Content Components: Teacher Self-Report	Sessions Offered: School Records and Teacher Self-Report	Attendance: School Records and Teacher Self-Report	Participant Engagement: Participant Self-Report	Participant Satisfaction: Participant Self-Report	Teacher Training: Study Records	Quality Summary: Score Calculated
Fidelity	X	X						
Dose delivered			X	X				
Dose received					X	X		
Program quality							X	X

Results measuring implementation fidelity found that although teachers scored well on their adherence to program protocol, there was large variation in the proportion of curriculum core content components covered by each group, ranging from 8% to 86%. Additionally, dose delivered also varied widely, with the number of sessions offered ranging from 7 to 46. Finally, despite high participant satisfaction, with 84% of students stating that they would recommend the program to others, there were large variations in the quality summary scores of program delivery. Overall, the evaluation findings reinforced the program, including enhancements to the curriculum, teacher training, and technical assistance. The evaluators noted the limitations of collecting self-reported data, but they also acknowledged the value of collecting data from multiple sources, allowing the triangulation of findings.

Source: Adapted from Morrel-Samuels et al. (2017).

Process evaluation is the most common form of program evaluation. It is used both as a stand-alone evaluation and in conjunction with impact assessment as part of a more comprehensive evaluation. As a stand-alone evaluation, it yields quality assurance information, assessing the extent to which a program is implemented as intended and operating according to the standards established for it. When the program model used is one of established effectiveness, establishing that the program is well implemented can be presumptive evidence that the expected outcomes are produced as well. When the program is new, a process evaluation provides valuable feedback to administrators and other stakeholders about progress implementing the program design. From a management perspective, process evaluation provides the feedback that allows a program to be managed for high performance, and the associated data collection and reporting of key indicators may be institutionalized in the form of a data dashboard to provide routine, ongoing feedback on key performance indicators.

In its other common application, process evaluation is an indispensable adjunct to impact assessment. The information about a program's effects on its target outcomes that evaluations of impact provide is incomplete and ambiguous

without knowledge of the program activities and services that produced those outcomes. When no impact is found, process evaluation has significant diagnostic value, indicating whether this was because of implementation failure, that is, the intended services were not provided hence the expected benefits could not have occurred, or theory failure, that is, the program was implemented as intended but failed to produce the expected effects. Process evaluation and program monitoring are described in more detail in [Chapter 4](#), and outcome monitoring is described in [Chapter 5](#).

Effectiveness of the Program: Impact Evaluation

The effectiveness of a social program is gauged by the change it produces in outcomes that represent the intended improvements in the social conditions it addresses. The ability of a program to have that impact will depend in large part on whether it adequately operationalizes and implements an effective theory of action grounded in an understanding of the social conditions in which it intervenes. **Impact evaluation** asks whether the desired outcomes were actually affected and whether the changes included unintended side effects. Examples of evaluation questions that might be addressed by impact evaluation include:

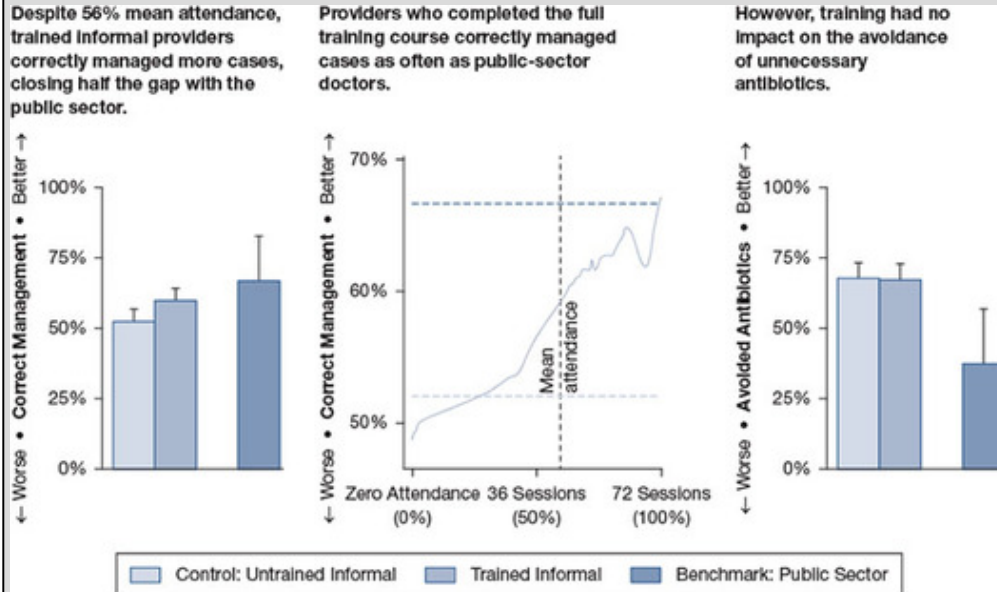
- Are the outcome goals and objectives of the program being achieved?
- Are the trends in outcomes moving in the desired direction?
- Does the program have beneficial effects on the recipients and what are those effects?
- Are there any adverse effects on the recipients, and what are they?
- Are some recipients affected for better or worse than others, and who are they?
- Is the problem or situation the program addresses made better? How much better?

The major difficulty in assessing the impact of a program is that the desired outcomes can usually also be influenced by factors unrelated to the program. Accordingly, impact assessment involves producing an estimate of the *net effects* of a program—the changes brought about by the intervention above and beyond those resulting from other processes and events affecting the targeted social conditions. To conduct an impact assessment, the evaluator must thus design a study capable of establishing the status of program recipients on relevant outcome measures and also estimate what their status would have been had they not received the intervention. Much of the complexity of impact assessment is associated with obtaining a valid estimate of the latter, known as the *counterfactual* because it describes a condition contrary to what actually happened to program recipients ([Exhibit 1-G](#) presents an example of an impact evaluation).

Exhibit 1-G Evaluating the Effects of Training Informal Health Care Providers in India

In many countries in the developing world, health care providers without formal medical training account for a large proportion of primary health care visits. Despite legal prohibitions in rural India, informal providers, who are estimated to exceed the number of trained physicians, provide up to three fourths of primary health care visits. Medical associations in India have taken the position that training informal providers may legitimize illegal practices and worsen public health outcomes, but there is little credible evidence on the benefits or adverse side effects of training informal providers. Because of the severe shortage of trained health care providers, an intervention to train informal health care providers was designed as stopgap measure to improve health care while reform of health care regulations and the public health care system was undertaken. The intervention took place in the Indian state of West Bengal and trained informal health care providers in 72 sessions over a period of 9 months on multiple topics, including basic medical conditions, triage, and the avoidance of harmful practices.

A randomized design was used to evaluate the impact of the training program. A sample of 304 providers who volunteered for the training was randomly split into treatment and control groups, the latter of which was offered the training program after the evaluation was complete. Daylong clinical observations that assessed the clinical practices of the providers and their treatment of unannounced standardized patients who were trained to present specific health conditions to the health care providers, were employed to test each participant on his or her delivery of treatment and utilization of skills taught in the training. The researchers withheld information about which group, treatment or control, the health care providers were in from the test patients. The researchers found that the training increased rates of correct case management by 14%, but the training had no effect on the use of unnecessary medicines and antibiotics. Overall, the results suggested that the intervention could offer an effective short-term strategy to improve health care provision. The graphic below provides a summary of the research results:



The evaluators raised concerns about the failure of the training to reduce prescriptions of unnecessary medications, even though it had been explicitly included in the training. They noted that many of the informal providers made a profit on the sale of prescriptions and stated, “We believe these null results are directly tied to the revenue model of informal providers.”

Source: Adapted from Das, Chowdhury, Hussam, and Banerjee (2016).

Determining when an impact assessment is appropriate and what evaluation design to use presents considerable challenges to the evaluator. Evaluation sponsors often believe they need an impact evaluation, and indeed, it is the only way to determine if the program is bringing about the intended changes. However, an impact assessment can be demanding of expertise, time, and resources and may be difficult to set up properly within the constraints of routine program operation. If the need for information about effects on outcomes is sufficient to justify an impact assessment, there is still a question of whether the program circumstances are suitable for conducting such an evaluation. For instance, it makes little sense to establish the impact of a program that is not well structured or cannot be adequately described. Impact assessment, therefore, is most appropriate for mature, stable programs with well-defined program models and a clear intention to use the results to justify the effort required. Impact assessment is also often appropriate for demonstration projects or pilots of programs that are under consideration for widespread adoption. [Chapters 6 to 8](#) discuss impact assessment and the various ways in which it can be designed and conducted.

Cost Analysis and Efficiency Assessment

Finding that a program has positive effects on the intended outcomes is often insufficient for assessing its social value. Resources for social programs are limited, so their accomplishments must also be judged against their costs. The first requirement for evaluations assessing costs is to describe the specific costs incurred in operating a program. Although many programs have expenditure records, the actual costs of operating a program may include donated items, volunteer time, and opportunity costs (costs associated with spending time on the program rather than other uses of the time by leaders, staff, and participants). A careful description of the full costs of a program is referred to as a **cost analysis**. Beyond describing the costs needed to operate a program, an **efficiency assessment** takes account of the relationship between a program's costs and its effectiveness. Efficiency assessments may take the form of **cost-benefit analysis** or **cost-effectiveness analysis**, asking, respectively, whether a program produces sufficient benefits in relation to its costs and whether other interventions or delivery systems can produce the benefits at a lower cost. Examples of evaluation questions that might guide an efficiency assessment are as follows:

- What are the actual total costs of operating the program, and who pays those costs?
- Are resources used efficiently without waste or excess?
- Is the cost reasonable in relation to the magnitude or monetary value of the benefits?
- Would alternative approaches yield equivalent benefits at less cost?

Exhibit 1-H Assessing the Cost-Effectiveness of Supported Employment for Individuals With Autism in England

In England, autism spectrum conditions affect approximately 1.1% of the population, and the costs of supporting adults with autism spectrum conditions is estimated to be £25 billion. Given that adults with autism experience difficulties in finding and retaining employment, and the employment rate for adults with autism is estimated to be 15%, the evaluators set out to estimate the cost-effectiveness of supported employment in comparison with standard care or day services.

The authors drew the data on program effectiveness from a prior evaluation, which found that a supported employment program specifically for individuals with autism in the United Kingdom increased employment and job retention in a follow-up study 7 to 8 years after the program was initiated. The program assessed the clients, supported them in obtaining jobs,

supported them in coping with the requirements for maintaining employment, educated employers, and advised coworkers and supervisors on how to avoid or handle any problems. For the main analysis, the evaluators used cost data from a study of the unit costs for supported employment services and day services for adults with mental health problems.

Table 1

Results of main analysis: mean costs, number of weeks in employment and QALYs of supported employment and standard care per adult with autism seeking employment, over the time horizon of the analysis (17 months of intervention + 8 years follow-up)					
	Mean total intervention cost over 17 months	Mean total day service cost over 8-year follow-up (incurred by unemployed only)	Mean total cost	Mean number of weeks in employment	Mean number of QALYs
Supported employment	£5044	£4193	£9237	136	5.42
Standard care (day services)	£2742	£5893	£8635	102	5.31
Difference	£2302	-£1700	£602	34	0.11
Cost-effectiveness	ICER of supported employment versus standard care: £18/extra week in employment; £5600/QALY				

QALY: quality-adjusted life year; ICER: incremental cost-effectiveness ratio.

Note that numbers have been rounded to the nearest £ (costs), to the nearest integer (weeks in employment) and to the nearest second decimal digit (QALYs).

QALY: quality-adjusted life year; ICER: incremental cost-effectiveness ratio.

Note that numbers have been rounded to the nearest £ (costs), to the nearest integer (weeks in employment) and to the nearest second decimal digit (QALYs).

The incremental cost-effectiveness analysis, or the cost of an extra week of employment, was £18, which led the authors to determine that supported employment programs for adults with autism were cost effective. The authors concluded, “Although the initial costs of such schemes are higher than standard care, these reduce over time, and ultimately supported employment results not only in individual gains in social integration and well-being but also in reductions of the economic burden to health and social services, the Exchequer and wider society.”

Source: Adapted from Mavranouzouli et al. (2014).

Efficiency assessment can be tricky and arguable because it requires making assumptions about the dollar value of program-related activities and, sometimes, imputing monetary value to program outcomes, both beneficial and adverse, that are difficult to represent with a dollar value. Nevertheless, such estimates are often germane for informing decisions about allocation of resources and identification of the program models that produce the strongest results with a given amount of funding. In certain cases, a descriptive cost analysis by itself may provide salient information to guide decisions about program adoption or consideration that involve fewer assumptions than efficiency assessments.

Like impact assessment, efficiency assessment is most appropriate for mature, stable programs with well-structured program models. This form of evaluation builds on process and impact assessment. A program must be well implemented and produce the desired effects before questions of how efficiently it accomplishes that become especially relevant. Given the specialized expertise required to conduct efficiency assessments, it is also apparent that it should be undertaken only when there is a clear need and identified use for the information. With the high level of concern about program costs in many contexts, however, this may not be an unusual circumstance. [Chapter 10](#) discusses cost and efficiency assessment methods in more detail.

The Interplay Among the Evaluation Domains

As is apparent in the descriptions above of the issues that motivate the different domains of evaluation questions, they reflect a general logic about what constitutes an effective program. That logic says that a program must correctly diagnose and understand the problem or conditions it aims to improve, be designed around a feasible plan for addressing the problem that is based on a valid theory about how the intended changes can be brought about, and operationalize that design in the way it is implemented and sustained. Those qualities should position the program to be effective, that is, to have a beneficial impact on the respective outcomes for the population targeted by the program. Being effective, however, does not necessarily mean being efficient. To be efficient, the program must achieve its effects at an acceptable cost to its sponsors and funders, and at a cost that compares favorably with other means of attaining the same effects.

There is a parallel logic for evaluators attempting to assess these various aspects of a program. Each family of questions draws on or makes assumptions about the answers to the prior questions. A program's theory and design, for instance, cannot be adequately assessed without some knowledge of the nature of the need the program is intended to address. If a program addresses lack of economic resources, the appropriate program concepts and the evaluation questions will be different than if the program addresses drunken driving. Moreover, the most appropriate criteria for judging program design and theory is how responsive it is to the nature of the need and the circumstances of those in need. When an evaluation of a program's theory and design are undertaken in the absence of a prior needs assessment, the evaluator must make assumptions about the extent to which the program design reflects the actual needs and circumstances of the target population to be served. There may be good reason to have confidence in those assumptions, but that will not always be the case.

Similarly, the central questions about program process are about whether the program operations and service delivery are consistent with the program theory and design; that is, whether the program *as intended* has actually been implemented. This means that the criteria for assessing the quality of the implementation are based, at least in part, on how the program is intended to function as specified by its basic conceptualization and design. The evaluator assessing program process must therefore be aware of the nature of the intended

implementation, perhaps from a prior assessment of the program theory and design, but more often by reviewing program documents, talking with key stakeholders, and the like. The quality of implementation for a program to feed the homeless through a soup kitchen cannot be assessed without knowing the aims of the program with regard to the population of homeless individuals targeted, the manner in which they are to be reached, the nature of the nutritional support to be provided, the number of individuals to be served, and other such specifics about the expectations and plans for the program.

Questions about program impact, in turn, are most meaningful and interpretable if the program is well implemented. Program services that are not actually delivered, are not fully or adequately delivered, or are not the intended services cannot generally be expected to produce the desired effects on the conditions the program is expected to impact. Evaluators call it **implementation failure** when the effects are null or weak because the program activities assumed necessary to bring about the desired improvements did not actually occur as intended. But a program may be well implemented and yet fail to achieve the desired impact because the program design and theory embodied in the corresponding program activities are faulty. When the program conceptualization and design are not capable of generating the desired outcomes no matter how well implemented, evaluators interpret the lack of impact as **theory failure**.

The results of an impact evaluation that does not find meaningful effects on the intended outcomes, therefore, are difficult to interpret when the program is not well implemented. The poor implementation may well explain the limited impact, and attaining and sustaining adequate implementation is a challenge for many programs. But it does not follow that better implementation would produce better outcomes; implementation failure and theory failure cannot be distinguished in that situation. Strong implementation, in contrast, allows the evaluator to draw inferences about the validity of the program theory, or lack thereof, according to whether the expected impacts occur. It is advisable, therefore, for the impact evaluator to obtain good information about program implementation along with the impact data.

Evaluation questions relating to program cost and efficiency also draw much of their significance from answers to prior evaluation questions. In particular, a program must have at least minimal impact on its intended outcomes before questions about the efficiency of attaining that impact become relevant to

decisions about the program. If there are no program effects, there is little for an efficiency evaluation to say except that any cost is too much.

Needs assessments, assessments of program theory and design, assessments of program process, impact evaluations, and cost analysis and efficiency assessments can all be conducted as stand-alone evaluation studies, and the questions addressed in each case will be meaningful in many program contexts. As we have shown, however, there is an interplay among these evaluation domains such that information about the issues addressed in each have implications for the questions, answers, and interpretations in other domains. Some of this can be thought of in relation to the life cycle of a program, with assessments of need, program theory, and program process ideally feeding successively into the planning and initial implementation of a new program. When full implementation is attained, impact evaluation can then test the expectation that this sequence has resulted in a program that has beneficial effects for its target population. If so, an efficiency assessment can guide consideration of whether the cost of achieving those benefits is acceptable. In the rough-and-tumble world of social programs, however, the need for actionable information from an evaluation will not always hew to this logic, and evaluations centered on any of the domains may be appropriate at different stages in the life cycle of a program.

Most of the remainder of this text is devoted to further describing the nature of the issues and methods associated with each of the five evaluation domains and their interrelationships.

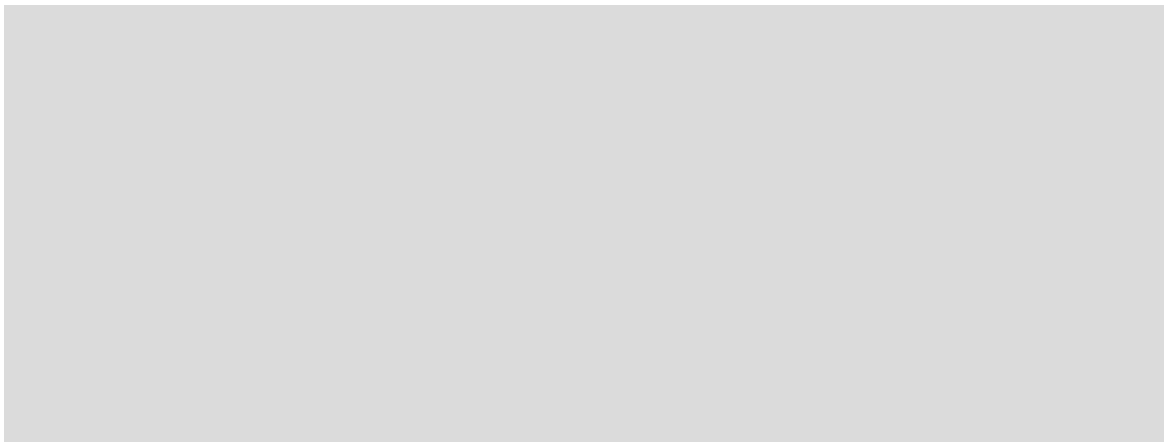
Summary

- Program evaluation focuses on social programs, especially human service programs, but the concepts and methods are broadly applicable to any organized social action.
- Most social programs are well intended and take reasonable approaches to improving the social conditions they address, but that is not sufficient to ensure they are effective; systematic evaluation is needed to objectively assess their performance.
- Program evaluation involves the application of social research methods to systematically investigate the performance of social intervention programs and inform social action.
- Evaluation has two distinct but closely related components, a description of performance and standards or criteria for judging that performance.
- Most evaluations are undertaken for one of three reasons: program improvement, accountability, or knowledge generation.
- The evaluation of a program involves answering questions about the program that generally fall into one or more of five domains: (a) the need for the program, (b) its theory and design, (c) its implementation and service delivery, (d) its outcome and impact, and (e) its costs. Each domain is characterized by distinctive questions along with concepts and methods appropriate for addressing those questions.

- Although program evaluations fall into one of these five domains, any particular evaluation involves working with key stakeholders to adapt the evaluation to its political and organizational context.
- Ultimately, evaluation is undertaken to support decision making and influence action, usually for the specific program that is being evaluated, but evaluations may also inform broader understanding and policy for a type of program.

Key Concepts

Assessment of program process 21
Assessment of program theory and design 19
Confirmation bias 5
Cost analysis 25
Cost-benefit analysis 25
Cost-effectiveness analysis 25
Demonstration program 10
Efficiency assessment 25
Empowerment evaluation 14
Evaluation questions 16
Evaluation sponsor 9
Formative evaluation 11
Impact evaluation 23
Implementation failure 28
Independent evaluation 13
Needs assessment 17
Outcome monitoring 21
Participatory or collaborative evaluation 14
Performance criterion 15
Program evaluation 6
Program monitoring 21
Social research methods 6
Stakeholders 9
Summative evaluation 11
Theory failure 28



Critical Thinking/Discussion Questions

1. Explain the four different reasons evaluations are conducted. How does the reason an evaluation is undertaken change how the evaluation is conducted?
2. Explain what is meant by systematic evaluation and discuss what is necessary to conduct an evaluation in a systematic way.
3. There are five domains of evaluation questions. Describe each of the five domains and discuss the purpose of each. Provide examples of questions from each of the five domains.

Application Exercises

1. At the beginning of the chapter the authors provide a few examples of social interventions that have been evaluated. Locate a report of an evaluation of a social intervention and prepare a brief (3- to 5-minute) summary of the social intervention that was evaluated and the evaluation that was conducted.
2. This chapter discusses the role of stakeholders, which are individuals, groups, or organizations with a significant interest in how well a program is working. Think of a social program you are familiar with. Make a list of all of the possible stakeholders for that program. How could their interest in the program be the same? How could they differ? Which stakeholders do you believe are most important to engage in the evaluation process and why?

Chapter 2 Social Problems and Assessing the Need for a Program

The Role of Evaluators in Diagnosing Social Conditions and Service Needs

Defining the Problem to Be Addressed

Specifying the Extent of the Problem: When, Where, and How Big?

Using Existing Data Sources to Develop Estimates and Identify Trends in Social Indicators

Estimating Problem Parameters Through Social Research

Agency Records

Surveys and Censuses

Key Informant Surveys

Forecasting Needs

Defining and Identifying the Target Populations of Interventions

Who or What Is a Target Population?

Specifying Targets

Target Boundaries

Varying Perspectives on Specification of the Target Population

Describing Target Populations

Risk, Need, and Demand

Incidence and Prevalence

Rates

Describing the Nature of Service Needs

Qualitative Methods for Describing Needs

Summary

Key Concepts

Understanding the nature of the social problem a program is intended to alleviate is fundamental to the evaluation of that program. The evaluation activities that examine the social problem are usually called needs assessment. From a program evaluation perspective, needs assessment is the means by which an evaluator determines whether there is a need for a program and, if so, the nature and extent of that need and related implications for program services appropriate to address that need. Needs assessment is critical for the design of new programs, but is also relevant for established programs

when it cannot be assumed that the program continues to meet the need or that the need has not changed.

Needs assessment is fundamental because a program cannot be effective at ameliorating a social problem if there is no problem to begin with or if it is not well enough understood to allow program services to be tailored in a way that is effective for addressing the problem. This chapter focuses on the role of evaluators in diagnosing social problems through systematic procedures in ways that can be related to the design and evaluation of programs.

As noted in [Chapter 1](#), effective programs are instruments for improving social conditions. In evaluating a social program, it is therefore essential to ask whether it addresses a significant social problem, and if it does so in a manner sufficiently responsive to the circumstances to plausibly bring about improvements. Answering these questions first requires a description of the social problem the program is designed to address. The evaluator can then ask whether the program's action theory embodies a valid conceptualization of the problem and an appropriate means of ameliorating it. If both questions are answered in the affirmative, the evaluator's attention can turn to whether the program is implemented in line with the program theory and, if so, whether the intended improvements in the social conditions actually result and at what cost. Thus, the logic of the different domains of program evaluation questions that was explained in [Chapter 1](#) builds fundamentally upon a description of the social problem the program addresses.

The procedures used by evaluators and other social researchers to systematically describe and diagnose social needs are referred to as **needs assessment**. The task for the evaluator as needs assessor is to describe the problem that concerns the relevant stakeholders in a manner that is as careful, objective, and meaningful as possible and to draw out the implications of that diagnosis for structuring an effective program. This task involves constructing a precise definition of the problem, assessing its scope and extent, identifying the **target population** for the program, and describing the characteristics of the problem and the target population that have implications for the design of services with the potential to respond effectively to the problem. This chapter describes these activities in detail.

The Role of Evaluators in Diagnosing Social Conditions and Service Needs

In their professional roles, evaluators are not major actors in the social and political advocacy process that identifies social problems and motivates organized efforts to address them. But effective responses to such problems often include intervention programs at the local grassroots, community, and national levels that must be based on careful documentation of the nature of the problems, with an emphasis on implications for effective programmatic solutions appropriate to the context. This is where evaluators make significant contributions by applying their repertoire of research techniques to systematically describe the problematic social conditions, gauge the appropriateness of proposed and established intervention programs, and assess the effectiveness of those programs for improving those conditions.

The importance of systematic information cannot be overstated. Speculation, impressionistic observations, political pressure, vested interests, and even deliberately biased information can shape the programs that policymakers, planners, and funders undertake or support in response to a perceived social problem. But if sound judgment is to be reached about such matters, it is essential that, to the extent possible, these actors have an adequate understanding of the nature and scope of the problem the program is meant to address, the relevant characteristics of the corresponding target population, and the context within which the program operates or will operate. Here are a few examples from the United States that illustrate what can happen when programs are based on inadequate description and diagnosis:

- Globalization has been touted as the main cause of the loss of manufacturing jobs in the United States as companies have moved their factories to Mexico, China, India, and other countries where cheaper labor and lower taxes could be found. Corporate tax rate reductions meant to offset the benefits of such offshoring have been enacted in part to keep relatively high paying, skilled labor employment in manufacturing in the United States. However, much of

the loss of manufacturing jobs is attributable to investments in technology that have reduced the number of workers required for the same level of productivity. Tax reductions may help keep manufacturing in the United States, but they do not address the substantial job loss stemming from increased use of robotics and other manufacturing technologies.

- Homelessness programs across the United States have largely designed their services for impoverished men, who often have substance abuse or mental health issues, and women with children who have fled domestic violence. A 2015 survey conducted in Atlanta, Georgia, identified a third subpopulation of more than 3,000 homeless youth, many members of the LGBTQ community (see [Exhibit 2-E](#)). Because the number of homeless youth had surged since prior efforts to survey the homeless, they were largely ignored by the public programs.
- Throughout the United States, most public health clinics offer family planning services that include subsidized birth control methods for poor women, many only teenagers. However, these clinics mainly provide access to cheaper and less effective forms of birth control, such as condoms or birth control pills, because of the higher cost of more effective, long-acting reversible contraceptives. Those cheaper forms of contraception, however, must be regularly replaced or renewed, which can be a challenge for economically stressed women. Intermittent access to and inconsistent use of these less effective forms of contraception have left many poor women and teenagers vulnerable to unwanted pregnancies.

In all these examples, a thorough needs assessment would have alerted policymakers and advocates to a broader set of needs, issues, and responses than recognized by existing policies and programs. More generally, a needs assessment can help keep unnecessary programs from being designed and implemented, help ensure that all the underlying conditions and subpopulations are addressed by the program, and help redirect program attention to changes in the respective social conditions, such as the emergence of a distinct subgroup of the target population that might otherwise be overlooked or incorrectly identified.

All social programs rest on assumptions about the nature of the problems they address and the characteristics, needs, and responses of the target populations they serve. Any evaluation of a plan for a new program, a change in an existing program, or the effectiveness of an ongoing program must necessarily consider those assumptions. Of course, these assumptions may already be supported by adequate evidence, in which case the evaluator can move forward on the basis of that evidence. Or the evaluation task may be stipulated in such a way that the nature of the need for the program is not a matter that requires investigation. Or program personnel and sponsors may believe they understand the social problems and target population so well that further inquiry is unnecessary. Such claims must be approached cautiously. It is remarkably easy for a program to be based on faulty assumptions, either through insufficient initial problem diagnosis, changes in the conditions or target population, or selective exposure or stereotypes that lead to distorted views of the nature of the intended beneficiaries and their life situations.

An evaluator should scrutinize the assumptions about the social problem and target population that shape the nature of a program. Where there is ambiguity, it may be advisable for the evaluator to work with key stakeholders to formulate those assumptions explicitly and conduct at least a minimal needs assessment to sharpen understanding of the problem addressed and the population served. For new program initiatives in the planning stage, or established programs whose utility has been called into question, it will often be appropriate to conduct a full-scale needs assessment.

It should be noted that needs assessment is not always done with reference to a specific social program or program proposal. The techniques of needs assessment are also used as planning tools and decision aids for policymakers who must prioritize among competing needs and claims. For instance, a regional United Way or a city council might commission a needs assessment to help determine the most critical or widespread needs in the community. Or a department of mental health might assess community needs for different mental health services so that resources can be distributed appropriately across its provider units. Although these broader comparative needs assessments are different in scope and purpose from

assessment of the need for a particular program, the applicable methods are much the same, and such assessments are generally conducted by evaluation researchers.

[Exhibit 2-A](#) provides an overview of the basic steps involved in a full-scale needs assessment. Note that some of these steps include significant involvement by stakeholders, something we highlighted in [Chapter 1](#) as an essential component of any evaluation. Useful book-length discussions of needs assessment applications and techniques can be found in Altschuld and Kumar (2010) and Watkins, Meiers, and Visser (2012).

Defining the Problem to Be Addressed

The question of what constitutes a social problem has occupied spiritual leaders, philosophers, and social scientists for centuries. Thorny issues in this domain revolve around what is meant by a need in contrast, say, to a want or desire, and what ideals or expectations should guide decisions by social actors about whether to intervene (cf. Watkins et al., 2012). For our purposes, the key point is that social problems are not objective phenomena. Rather, they are social constructions involving assertions that certain conditions constitute problems that require public attention and deliberate, organized intervention. In this sense, community members, together with the stakeholders involved in a particular issue, create the social reality within which a social problem is defined (Miller & Holstein, 1993; Spector & Kitsuse, 1977).

It is generally agreed, for example, that poverty is a social problem. The observable facts are statistics on the distribution of income and assets. However, those statistics do not define poverty, they merely permit one to determine how many people are poor when a definition is given. Nor do they establish poverty as a social problem; they only characterize a situation that individuals and social agents may view as problematic. Moreover, both the definition of poverty and the goals of programs to improve the lot of the poor can vary over time, among communities, and among stakeholders. Initiatives to reduce poverty, therefore, may range widely, for example, from increasing employment opportunities to simply providing cash benefits to persons with low income. In [Exhibit 2-B](#), we see how the federal definition of poverty varies by family composition and size.

Exhibit 2-A The Three Phases of Needs Assessments

Phase 1: Preassessment

Purpose: Getting organized and establishing what is known

A. Scope out the problem in an exploratory fashion

What is the problem? Who is affected? What is currently being done? What seems to be causing the problem?

B. Identify key stakeholders and form a needs assessment work group or committee

Which groups or individuals are interested? Are their existing organizations focused on this problem or actively working on solving it? Are there political agendas that might be negatively affected?

C. Define the gap between the desired outcomes (what should be) and existing conditions (what is) on an initial, preliminary basis

Where can information about the problem be readily found? Are there existing reports, evaluations, or databases on the problem, who is affected, and the services currently offered?

D. Synthesize and communicate the evidence

What does the existing, readily available evidence say about the problem, who is affected, and what's being done? How can this be meaningfully and succinctly conveyed to the stakeholders?

E. Decisions and next steps

Are the needs, their importance, and the risks involved sufficiently well understood to make decisions? If not, what additional information is needed to make decisions?

Phase 2: Assessment

Purpose: Conduct a needs assessment; develop a deeper understanding of the needs of the target population, service deliverers, and organization(s) responsible; identify likely causes of the needs

F. Orientation of the assessment

What are the gaps between “what is” and “what should be” for the target population, service providers, and organizations responsible? Who is affected? What additional information is needed? What are the criteria for choosing a solution? What resources are needed for the assessment and how might they be obtained?

G. Plan for data collection

What data are needed? Will the data be collected through surveys, interviews, focus groups, or existing sources? How will the data be analyzed? How will the quantitative and qualitative findings be synthesized? How will the synthesis be communicated?

H. Collect, analyze, and synthesize data

Obtain appropriate approvals to collect data from human subjects. Determine from whom the data will be collected. Administer surveys to study samples. Schedule and conduct interviews and focus groups. Obtain the existing data from secondary sources. Analyze the data from each data source. Synthesize and summarize the data across sources and triangulate the evidence.

I. Decisions and next steps

Meet with the group of key stakeholders. Discuss potential benefits, risks, and potential adverse consequences for each potential remedy.

Phase 3: Postassessment

Purpose: Take action to resolve the problem underlying the need

J. Review and reconstitute the key stakeholder group

What organizations are involved to address needs? Have the needs of the target population, service providers, and organizations been identified and considered? Are all of the key stakeholders and representatives of the organizations currently involved in the process? What criteria will be used to determine which remedy will be selected?

K. Analyze potential causes of the needs and remedies for the gaps

What are the likely causes of the gaps that have been prioritized? Which potential remedies are considered most likely to eliminate or ameliorate the needs? How do the potential remedies rate on the criteria for choosing a remedy?

L. Select the solution to be implemented

Determine the ranking of each potential solution on the basis of the criteria for choosing a remedy. Select the remedy. Develop an action plan for implementing the remedy. Obtain resources to implement the remedy. Implement the remedy and monitor the process and outcomes. Evaluate the remedy.

Source: Adapted from Altschuld and Kumar (2010).

Exhibit 2-B Federal Definition of Poverty



Defining a social problem and specifying the goals of intervention are thus ultimately social and political processes that do not follow automatically from the inherent characteristics of the situation. This circumstance is illustrated nicely in an analysis of legislation designed to reduce adolescent pregnancy conducted by the U.S. General Accounting Office (1986), which found that none of the pending legislative proposals defined the problem as involving the fathers of the children in question; each addressed adolescent pregnancy only as an issue of young mothers. Although this view of adolescent pregnancy may lead to effective programs, it nonetheless clearly represents arguable assumptions about the nature of the problem and how a solution should be approached.

The social definition of a problem is so central to the political response that the preamble to proposed legislation usually includes some attempt to specify the problematic condition the legislation is designed to remedy. For example, two contending legislative proposals may be addressed to the problem of childhood obesity, but one may describe that problem as one of poor food choices by parents and children, whereas the other may identify it as one of limited access to nutritional alternatives (e.g., food deserts in poor neighborhoods and the prevalence of sugary drinks in schools). The first perspective centers attention primarily on the personal behavior of the target population; the second focuses on the environments within which that population lives. The ameliorative actions justified by these perspectives will be different as well; the first suggests programs to educate children and parents about healthy foods and obesity, the second would support

restrictions and incentives to provide greater access to nutritious food and drink.

It is usually informative, therefore, for an evaluator to determine what the major political actors think the problem is. In the preassessment phase in [Exhibit 2-A](#), the evaluator might, for instance, study the definitions given in policy and program proposals or enabling legislation. Such information may also be found in legislative proceedings, program documents, newspaper and magazine articles, and other sources in which the problem or program is discussed. Such materials may explicitly describe the nature of the problem and the program's plan of attack, as in funding proposals, or implicitly define the problem through the assumptions that underlie statements about program activities, successes, and plans. This inquiry will almost certainly turn up information useful for a preliminary description of the social need to which the program is expected to respond. As such, it can guide a more probing needs assessment with regard to both how the problem is defined and the alternative perspectives that might be applicable.

Specifying the Extent of the Problem: When, Where, and How Big?

Having defined the problem a program addresses, an evaluator can then assess the scope and extent of that problem. Ideally, the design and funding of a social program would be geared to the size, distribution, and density of the problematic condition. In assessing, say, emergency shelters for victims of domestic violence, it makes a difference whether the number of victims seeking shelter in the community at any one time is 50 or 500. It also matters whether such victims are primarily located in urban, suburban, or rural areas, and how many have children, close relatives, injuries requiring medical attention, or are employed.

It is much easier to establish that a problem exists than to develop valid estimates of its density and distribution. Identifying a handful of battered children may be enough to convince a skeptic that child abuse exists. But specifying the size of the problem and where it is located geographically and socially requires detailed knowledge about the population of abused children, the characteristics of the perpetrators, and where both are located. For a problem such as child abuse, which is not generally public behavior, this can be difficult. Many social problems are mostly invisible from any public perspective, so that only imprecise estimates of their rates are possible. In such cases, it is often necessary to use data from several sources and apply different approaches to estimating incidence rates.

It is also important to have at least reasonably representative samples with which to estimate the extent of a problem. It can be especially misleading to draw estimates from samples such as participants in the service programs that serve the target population, who are likely different in many ways from the population as a whole. Estimation of the rate of spousal abuse during pregnancy on the basis of reports from residents of battered women's shelters, for instance, will result in overestimation of the frequency of occurrence in the general population of pregnant women. Probability sampling is a commonly used method in the social sciences to ensure that

the characteristics of a sample of the population can be used to estimate the characteristics of the full population from which the sample was drawn.

The definition of a **probability sample** is that every member of the target population has a known, nonzero chance of being selected for the sample. This means that selection into the sample is done randomly; in other words, selection is a matter of chance that eliminates systematic bias in the selection process. Key to obtaining a probability sample is the availability of a complete list of the units in the target population to use as the **sampling frame**. A good sampling frame lists (a) all units (usually individuals) that are members of the target population, (b) no units that are not members of the target population, (c) no units more than once (duplicate entries), and (d) individual units rather than groups or clusters of units (individuals rather than households, for example).

If a sampling frame lacks any of these desirable characteristics, it may be possible to address that discrepancy in a more sophisticated sampling design. For instance, if an available list of members of the target population omits some members but the evaluator has a complete list of clusters, such as treatment facilities or schools that contain all members of the target population, these clusters can be sampled rather than the members individually. Data are then collected from all the eligible units within each of the sampled clusters. An overview of sampling designs is provided in [Exhibit 2-C](#), and book-length treatments (e.g., Henry, 1990; Kish, 1995) provide further detail. The surveys conducted by the U.S. government that are described in the next section of this chapter rely on probability sampling, usually multistage sampling, to reduce bias and increase the representativeness of the data.

Using Existing Data Sources to Develop Estimates and Identify Trends in Social Indicators

For some social issues, existing data sources, such as administrative data, surveys, and censuses, may be of sufficient quality to be used with confidence for assessing certain aspects of a social problem. For example, accurate and trustworthy information can usually be obtained from data collected by the American Community Survey of the U.S. Census Bureau or the decennial U.S. census. The decennial census reports data on census tracts (small areas containing about 4,000 households) that can be aggregated to get neighborhood and community data. When evaluators use sources whose validity is not as widely recognized as that of the census, they must assess the validity of the data by examining carefully how they were collected. A good rule of thumb is to anticipate that, on any issue, different data sources may provide disparate or even contradictory estimates.

On some topics, existing data sources provide periodic measures that chart historical trends. For example, the Current Population Survey of the Census Bureau collects annual data on the characteristics of the U.S. population from a large household sample. These data include composition of households, individual and household income, and household members' age, sex, and race. The regular Survey of Income and Program Participation provides data on U.S. population participation in various social programs, such as unemployment benefits, disability income, health insurance, income assistance, food benefits, job training programs, and so on.

A regularly occurring measure such as those mentioned above, called a **social indicator**, can provide especially useful information for assessing social problems and needs. First, these data can often be used to estimate the size and distribution of the social problem whose course is being tracked over time. Second, the trends shown can be used to alert decision makers to whether the pertinent conditions are improving, remaining the same, or deteriorating. As an example, [Exhibit 2-D](#) describes the use of American Community Survey data to describe the population of children

living in poverty in New Orleans and to explore possible reasons for their poverty status. This needs assessment showed that two thirds of single mothers living in poverty were working, but in low-wage jobs, suggesting that better jobs for parents and supportive services for their children may be needed to ameliorate the effects of poverty on children's development.

Exhibit 2-C Probability Sampling Designs

Probability Sampling Design	Definition	Salient Characteristics
Simple random sample	All members of the study population have an equal chance of selection in the sample.	- Simple to conduct. - Requires a complete list of the study population.
Systematic sample	All members of the target population have a known, nonzero chance of selection in the sample.	- Requires an interval (i) at which sample units are selected ($i = N/n$) and random start between 1 and i. - Can be used in situations in which a complete list of the target population is not available but partial lists (or actual members of the target population) can be accessed at different sites.
Cluster sample	- All units of the target population are members of one and only one cluster. - All clusters have an equal chance of selection into the sample - Data are collected on all units in the randomly selected clusters.	- Often used when members of the target population are in naturally occurring groups, such as teachers in schools or case workers in county social service offices. - Reduces precision compared with a simple random sample of the same size (amount of reduction depends on how much of the variation of the variable of interest is between clusters; more within-cluster variation improves precision).
Stratified random sample	- All units of the target population are placed into one and only one stratum. - A known, nonzero sample of units is selected from each stratum. - The probability of selection within strata can be equal or unequal depending on study goals.	- Often useful when smaller subpopulations that are proportionately larger in some strata are of particular interest for the evaluation (e.g., underrepresented groups). - Improves precision compared with a simple random sample of the same size (amount of improvement depends on the correlation between strata assignment variable and variable of interest).
Multistage sample	- Similar to cluster samples in that all members of the target population are members of a cluster. - Clusters are sampled at the first stage, and then units are selected in the second stage for two-stage samples. - Multiple clustering such as individuals within census tracts within counties can be combined. - Clusters can be stratified before sampling.	- Often useful in large, multipurpose probability samples drawn at the international, national, or state or provincial level. - Requires sampling expertise to compute sampling weights if the probability of selection for final units is unequal and to calculate standard errors for hypothesis testing.

Source: Adapted from Henry (1990).

Source: Adapted from Henry (1990).

Exhibit 2-D Using Census Data to Assess the Needs of Children Living in Poverty

Motivated by the fact that the child poverty rate for New Orleans had climbed to 39% in 2013, nearly equaling the rate before the devastation from Hurricane Katrina, the Data Center for Southeastern Louisiana analyzed Census Bureau data from the American Community Survey and other data sources to describe the population of children living in poverty in New Orleans. The American Community Survey is an annual survey of 3.5 million households that gathers data on numerous employment, housing, and family variables.

First, child poverty in New Orleans was described with the following results:

Child poverty: The child poverty rate in New Orleans was the ninth highest for midsized cities in the nation, lower than Cleveland's rate but significantly higher than many comparable cities in the southeastern United States, including Tampa, Raleigh, and Virginia Beach.

Child poverty trends: The child poverty rate declined from 41% in 1999 to 32% following Katrina in 2007, then reversed, growing to 39% in 2013.

Family structure and child poverty: In midsized cities, the child poverty rate is negatively correlated with the percentage of children living with married parents.

Family structure and household poverty: The poverty rate for single-mother households in New Orleans increased from 52% in 1999 to 58% in 2013.

Female-headed households in poverty and employment: Despite high poverty rates, 67% of the single mothers in New Orleans were employed.

Prevalence of low-wage jobs in New Orleans: Twelve percent of full-time, year-round workers in New Orleans earned less than \$17,500 per year, compared with 8% nationally.

When these findings were combined with additional data, the needs assessment concluded:

Given the cost of living in New Orleans, a single worker needs a wage of roughly \$22 per hour to adequately provide for one child.

Research shows that child poverty can create chronic, toxic stress that leads to difficulties in learning, memory, and self-regulation.

Innovation will be required to break the cycle of poverty that threatens the development of children in New Orleans.

Two-generation approaches to give children access to a high-quality early childhood education, while helping parents get better jobs and build stronger families, may be required to ameliorate the effects of child poverty.

Source: Adapted from Mack (2015).

Estimating Problem Parameters Through Social Research

In many instances, no existing data source will provide estimates of the extent and distribution of a problem of interest. For example, there are no ready sources of information about household pesticide misuse that would indicate whether it is a problem in households with children. In other instances, good information about a problem may be available for a national or regional sample that cannot be disaggregated to a relevant local level. The National Survey of Household Drug Use, for instance, uses a nationally representative sample to track the nature and extent of substance abuse. However, the number of respondents from most states is not large enough to provide good state-level estimates of drug abuse, and no valid city-level estimates can be derived at all.

When pertinent data are nonexistent or insufficient, the evaluator must consider collecting new data. There are various ways to obtain relevant data, ranging from expert opinion to large-scale **sample surveys**. Decisions about the research effort to undertake must be based in part on the resources available and how important it is to have precise estimates. If, for legislative or program design purposes, it is critical to know rather exactly, say, the number of obese teenagers in a political jurisdiction, a carefully planned household survey may be necessary. In contrast, if the need is simply to determine whether teenage obesity exists in the jurisdiction, input from knowledgeable informants may suffice. Three types of data sources from which evaluators can obtain pertinent data are described below.

Agency Records

Information contained in the records of organizations that provide services to the population in question can be useful for estimating the extent of a social problem (Hatry, 2015). Some agencies keep excellent records on their clients, although others do not. When an agency's clients include all the persons manifesting the problem in question and records are faithfully kept, the evaluator may not need to search further. Unfortunately, these conditions are rather rare. For example, an evaluator may hope to be able to

estimate the extent of drug abuse in a certain locality by extrapolating from the records of persons treated in drug treatment clinics. To the extent that the local drug-using population participates fully in those clinics, such an estimate may be accurate. However, if all drug abusers are not served by those clinics, which is more likely, the prevalence of drug abuse will be more widespread than such an estimate would indicate.

Surveys and Censuses

When it is necessary to get very accurate information on the extent and distribution of a social problem and there are no existing credible data, the evaluator may need to undertake original research using sample surveys or censuses (complete enumerations). Either of these approaches can require considerable effort and technical skill as well as a substantial commitment of resources. To illustrate one extreme, [Exhibit 2-E](#) describes the needs assessment survey undertaken to estimate the size and composition of the homeless youth population in Atlanta. Although there was ample evidence that large numbers of youth were living on the streets, no reliable information was available about either the size of this population or the reasons for their homelessness. Researchers from local universities therefore undertook a survey to provide that information.

Exhibit 2-E A Survey Using Capture-Recapture to Study Homeless Youth

Each year, federal and state officials develop point-in-time estimates of the homeless population in the United States by conducting a survey of the sheltered and unsheltered homeless populations on a single night in January. However, this methodology may not be adequate for hard-to-reach populations, such as homeless youth. Wright et al. (2016) used systematic capture-recapture methods to accurately describe the current population of homeless youth in metropolitan Atlanta. Capture-recapture methods, originally developed to estimate the size of wildlife populations, have also been used to estimate the size of hard-to-find populations such as persons involved in criminal activity, drug use, and high-risk health behaviors (Bloor, Leyland, Barnard, & McKeganey, 1991; Rossmo & Routledge, 1990; Smit, Toet, & van der Heijden, 1997).

Wright et al. (2016) enlisted the help of community outreach teams that routinely work with homeless populations to implement a two-sample capture-recapture survey. Before the survey period, the outreach teams distributed LED keychain flashlights (a “capture token”) to the homeless youth they encountered during their regular activities. These flashlights were fluorescently colored so as to be memorable to anyone who saw them, and the outreach teams were instructed to show them to each homeless youth even if it

was not accepted as a gift. Any homeless youth who saw the flashlights during this period was “captured” for the purposes of the study. During the survey period that followed, any homeless youth encountered were asked whether they had seen the flashlight offered by the outreach teams during the prior weeks. Participant who remembered seeing the flashlight were coded as “recaptured.”

Using statistical algorithms based on the recapture rate, the researchers were able to estimate that there were approximately 3,374 homeless youth in any given summer month. This estimate was substantially larger than most governmental and community homeless service providers previously believed. Furthermore, estimates from capture-recapture sweeps made at different times revealed rapid social mobility for this population.

Source: Adapted from Wright et al. (2016).

Needs assessment surveys are typically more straightforward than the capture-recapture procedure described in [Exhibit 2-E](#). Often conventional sample surveys can provide adequate information. Unfortunately, the method for gathering survey data most common in the past, telephone interviews, has become outmoded because of the growing prevalence of cell phones and their use to screen calls. Currently, the main methods for administering surveys are by mail, face to face, or on the Internet with requests for participation sent via e-mail. Mail surveys for community needs assessment can often be conducted by using household lists from utility services, such as water or electrical providers. [Exhibit 2-F](#) offers an example of a successful census survey on community needs that was conducted in a small town in Nebraska. Households were identified from utility billing records, and surveys were distributed by mail and by volunteers.

Exhibit 2-F Assessing Community Needs Through Surveys

To gauge the needs among residents of a small town in Nebraska, the South Central Economic Development District developed and administered a census survey of residents in 2014. Using an address list based on utility billing information, surveys were distributed to households by volunteers or mail. Responses were received from 773 of the 998 households, for a 77% response rate. The survey collected data on numerous community issues, with some of the responses summarized below.

Seventy-eight percent of the respondents supported developing a plan to include new residential areas into the city boundaries.

Utility services, the City Park, and law enforcement services were rated good, but control of loose pets was rated only as fair.

Two thirds of the residents considered residential streets to be good, while two thirds considered the condition of the sidewalks to be poor or fair. Only 16% supported paying for sidewalk improvements through an assessment. Among several types of community projects a majority supported hiking and biking trails and paving gravel roads. Fifty-five percent of the households using child care indicated that quality care was very difficult to find, and another 36% indicated it was at least somewhat difficult to find. The survey identified strengths and challenges for this community in several categories, such as overall community quality of life, recreational facilities, education, child care, housing, and business development.

Source: Adapted from Hueftle (2014).

Many survey organizations have the capability to plan, carry out, and analyze sample surveys for needs assessment. In addition, it is often possible to add questions to regularly conducted studies in which different organizations buy time, thereby reducing costs. Whatever the approach, it must be recognized that designing and implementing sample surveys can be a complicated endeavor requiring high skill levels. For many evaluators, the most sensible approach may be to contract with a reputable survey organization for such work. For further discussion of the various aspects of sample survey methodology, see Fowler (2014) and Dillman, Smyth, and Christian (2014).

Exhibit 2-G Key Informant Identification of Public Health Priorities

The Milwaukee Health Care Partnership interviewed 41 key informants about the public health priorities for Milwaukee County, Wisconsin. The selected informants included representatives from city and county health agencies, advocacy organizations with interests in public health issues, local philanthropic organizations, hospitals and medical colleges, community service organizations, and city councils, among others.

Each informant was asked to rank up to five public health issues he or she considered most important for the county. For each of those issues, informants were then asked to comment on (a) existing strategies to address the issue, (b) barriers and challenges to addressing the issue, (c) additional strategies needed, and (d) key groups in the community that health services should partner with to improve community health.

The top priority public health issues identified by these informants were

- behavioral health, especially mental health and alcohol and drug issues;
- access to health care services;
- physical activity, obesity, and nutrition;
- health insurance coverage; and
- infant mortality.

Among these, mental health was the issue most often identified by the key informants as needing significant change and community investment. The barriers and challenges they highlighted included stigma and lack of general knowledge about mental health, issues within the service system (e.g., reimbursement, lack of providers, and lack of preventive services), unemployment and poverty, lack of Spanish-speaking and Latino providers, cost of care, transportation for patients, lack of education and training for public sector employees, a siloed system of organizations and providers, and lack of funding for needed programs.

The strategies most often mentioned by informants for addressing these barriers and challenges included devoting additional funds and providers to mental health issues, expanding health care coverage and age- and culturally appropriate programs (especially for Latinos), increasing mental health awareness, providing screening, and education starting in schools and continuing throughout the life course, integrating mental health into primary care settings, and reimbursing supporting care agencies.

More broadly, the key informants believed that community education for the general public and professionals could increase understanding of and compassion for individuals struggling with mental health issues. They also suggested improving care management and coordination across the community, a greater focus on holistic health, and working toward a community system of care that integrates services and providers.

Source: Adapted from Kessler (2013).

Key Informant Surveys

Perhaps the easiest, though by no means most reliable, approach to estimating the extent of a social problem is to ask **key informants**: persons whose position or experience gives them some knowledge of the nature, magnitude, and distribution of the problem at issue. Key informants can often provide useful information about the characteristics of a target populations and the nature of service needs (see [Exhibit 2-G](#) for an example). However, few informants have a vantage point or information sources that permit good estimation of the actual number of persons affected by a social problem, or the demographic and geographic distribution of those persons.

Although key informant input has limitations, it is relatively easy to obtain and can provide insights unavailable from other sources. Nonetheless, the information from key informant surveys must be viewed cautiously, given the potential for error and the potential for inconsistent reports from different informants. As illustrated by [Exhibit 2-G](#), there are topics on which key informants can provide useful information and important

insights. In all cases, the evaluator should choose informants who have appropriate expertise and ensure that they are questioned in a careful manner, including probing for the experiences or evidence they are drawing on when they respond.

Forecasting Needs

Both in formulating policies and programs and in evaluating them, it is often necessary to estimate what the magnitude of a social problem is likely to be in the future. A problem that is serious now may become more or less serious in later years, and program planning must attempt to take such trends into account. Yet the forecasting of future trends can be quite risky, especially as the time horizon lengthens.

There are a number of technical and practical difficulties in forecasting that derive in part from necessary assumptions about how the future will be related to the present and past. For example, at first blush a projection of the number of persons in a population who will be 18 to 30 years of age a decade from now seems easy to construct from the age structure in current population data. However, had demographers made such forecasts years ago for central Africa, they would have been substantially off the mark because of the unanticipated and tragic impact of the AIDS epidemic on young adults. Projections with longer time horizons would be even more problematic because they would have to take into account trends in fertility, migration, and mortality.

We are not arguing against the use of forecasts in needs assessment. Rather, we only caution against accepting forecasts uncritically without a thorough examination of how they were produced and recognition of any self-interest or political agendas by the organizations that produced them. For simple extrapolations of existing trends, the assumptions on which a forecast is based may be easily ascertained. For sophisticated projections such as those developed from multiple-equation, computer-based models, examining the assumptions may require the skills of an advanced programmer and an experienced statistician. Evaluators must recognize that all but the simplest forecasts are technical activities that require specialized knowledge and procedures and, at best, involve inherent uncertainties.

Defining and Identifying the Target Populations of Interventions

For a program to be effective, those implementing it must not only know what its target population is but also be able to readily direct its services to that population and screen out individuals who are not part of that population. Consequently, delivering service to a target population requires that the definition of the target population permit eligible individuals to be distinguished from those ineligible for program participation in a relatively unambiguous and efficient manner.

Specifying a program's target population is complicated by the fact that the definition of the population and its size may change over time. For instance, the populations of individuals with substance abuse disorders historically have consisted chiefly of users of such illegal drugs as heroin and cocaine. In recent years, however, there has been a large upsurge in the abuse of prescription drugs, especially opioids, which has significantly changed the nature of the target populations for drug treatment programs.

Who or What Is a Target Population?

The target population of a social program usually consists of individuals. But populations also may be groups (families, work teams, organizations), geographically and politically related areas (such as communities), or physical units (houses, road systems, factories). It is important at the outset of a needs assessment to clearly define the units that constitute the target population. For individuals, the target population is usually identified in terms of its members' social and demographic characteristics or their problems, difficulties, and conditions. Thus, targets of an educational program may be designated as children aged 10 to 14 who are 1 to 3 years below their normal grades in school. The targets of a maternal and infant care program may be defined as pregnant women and mothers of infants with annual incomes less than 150% of the poverty level.

When aggregates (groups or organizations) are members of a target population, they may be defined in terms of the characteristics of the individuals who constitute them (e.g., their collective properties and shared problems). For example, an organizational-level target for a prekindergarten improvement intervention might be centers or schools providing educational and child care services to 4-year-olds with at least 10 children enrolled. Some aggregate units, on the other hand, do not involve any reference to the individuals in the aggregate. A weatherization program for houses built before modern insulation techniques were the norm, for instance, involves a target population of houses whose age and physical characteristics are the defining features.

Another criterion for defining the target population is geographic. Needs assessments are geographically bounded, often by a political jurisdiction such as a county or province, or a region that may be a neighborhood or an established community. The geographic boundary for a needs assessment should be resolved in the preassessment phase ([Exhibit 2-A](#)) and examined critically through interactions with key stakeholders. In addition to governmental boundaries, there may be service-delivery boundaries that define the catchment area of the program for which a needs assessment is being done. Although setting the geographic boundaries for a needs

assessment may seem straightforward, the complexity of reaching a consensus among stakeholders and ensuring that the definition will be verifiable when deciding eligibility for the program can present challenges.

A further distinction is often relevant to the definition of a target population. In many cases the program that serves or is being planned to serve that target population has eligibility requirements that constrain who can receive services. For example, eligible recipients may need to qualify on the basis of low income or a defined risk for an adverse outcome. Such programs are referred to as **targeted programs**. In contrast, **universal programs** are open to broad target populations with few or no constraints (e.g., programs in public parks open to all who wish to participate, afterschool programs that accept any child in the school district parents wish to enroll).

Target populations may also be regarded as direct or indirect, depending on whether services are delivered to them directly by the programs or indirectly through activities the programs arrange. Most programs specify direct targets, as when a medical intervention treats persons with a given illness. However, in some cases, for either economic or feasibility reasons, programs may be designed to affect a target population by acting on an intermediary population or condition that will, in turn, have an impact on the intended target population. A rural development project, for example, might select influential farmers for intensive training with the expectation that they will persuasively share what they have learned with other farmers in their vicinity who, thus, are the indirect targets of the program. Similarly, professional development may be provided to teachers with the intent of improving their classroom practices in ways that result in greater student achievement.

Specifying Targets

At first glance, specifying the target population for a program may seem simple. However, although target definitions are easy to write, the results often fall short when the program or the evaluator attempts to use them to identify who is properly included or excluded from program services. There are few social problems that can be easily and convincingly described in terms of simple, unambiguous characteristics of the individuals experiencing the problem.

What, for instance, is a resident with cancer in a given community? The answer depends on the meanings of both “resident” and “cancer.” Does “resident” include only permanent residents, or does it also include temporary ones (a decision that would be especially important in a community with a large number of vacationers). As for “cancer,” are patients currently in remission included, and, whether they are or not, how long without a relapse constitutes recovery? Are cases of cancer defined only as diagnosed cases, or do they also include persons whose cancer had not yet been detected? Are all cancers included regardless of type or severity? Although it should be possible to formulate answers to questions such as these for a given program, this illustration shows that it may not be a simple matter for an evaluator to determine exactly how a program’s target population is defined.

Target Boundaries

Adequate specification of a target population establishes boundaries, that is, rules determining who or what is included and excluded. One risk in specifying target populations is a definition that is overinclusive. For example, specifying that a criminal is anyone who has ever violated a law is uselessly broad; only saints have not at one time or another violated some law, wittingly or otherwise. This definition is too inclusive, lumping together in one category trivial and serious offenses and infrequent violators with habitual felons.

Definitions may also prove too restrictive or narrow, sometimes to the point that almost no one falls into the target population. Suppose that the designers of a program to rehabilitate released felons decide to include only those who have never been drug or alcohol abusers. The extent of prior substance abuse is so large among released prisoners that few would be eligible given this exclusion. In addition, because persons with longer arrest and conviction histories are more likely to be past or current substance abusers, this definition eliminates those most in need of rehabilitation as eligible for the proposed program.

Useful target definitions must also be feasible to apply. A specification that hinges on characteristics that are difficult to observe or for which existing records contain no data may be virtually impossible to put into practice. Consider, for example, the difficulty of identifying individuals eligible for a job training program if they are defined as persons who hold favorable attitudes toward accepting help of that sort. Complex definitions requiring detailed information may be similarly difficult to apply. The data required to identify a target population of “former members of producers’ cooperatives who have planted barley for at least two seasons and have adolescent sons” would be difficult, if not impossible, to gather. In some cases, the definitions can be so cumbersome to apply, especially when reestablishing eligibility is required on a frequent basis, that the bureaucratic process required to prove eligibility can inhibit program participation and limit the benefits that could have resulted.

Varying Perspectives on Specification of the Target Population

Another issue in the definition of target populations can arise from differing perspectives by professionals, politicians, and other stakeholders involved—including, of course, the potential recipients of services. Discrepancies may exist, for instance, among the views of legislators at different levels of government. At the federal level, Congress may plan a program to alleviate the financial burden of natural disasters for a target population viewed as residents of areas in which 100-year floods may occur. True to their name, however, 100-year floods occur in any one place rather infrequently. From a

local perspective, individuals living in a flood plain that has not experienced flooding for many decades may not be viewed as a part of the target population, especially if it means that the local government must implement expensive flood control measures.

Similarly, differences in perspective can arise between program sponsors and the intended beneficiaries. The planners of a program to improve the quality of housing available to poor persons may have a conception of housing quality much different from that of the people who live in those dwellings. The program's definition of what constitutes the target population of substandard housing for renewal, therefore, could be much broader than what the residents of those dwellings view as adequate housing.

Although needs assessment cannot establish which perspective on a program's target population is correct, it can help eliminate conflicts that might arise from groups talking past one another. To accomplish this, evaluators should elicit the perspectives of all the significant stakeholders and ensure that none of those with a stake in the program are left out of the decision process through which the target population is defined. In this endeavor, evaluators may strive to meet the criteria for democratic process, which are inclusion, dialogue, and deliberation (House & Howe, 1999). These authors suggest that evaluators find ways to include all stakeholders, including those potentially eligible for program participation, in a genuine exchange of views while minimizing the political imbalances among the groups.

Describing Target Populations

The nature of the target population a program attempts to serve naturally has considerable implications for the program's approach and likelihood of success. In this section, we discuss a range of concepts useful for describing target populations in ways that highlight those implications.

Risk, Need, and Demand

A public health concept, **population at risk**, is helpful in specifying eligibility for interventions that address conditions that have not yet been experienced. The population at risk consists of those persons or units with a significant probability of experiencing or developing the condition to which the program is designed to respond. Thus, the population at risk in birth control programs is usually defined as women of childbearing age. Similarly, projects designed to mitigate the effects of typhoons and hurricanes may define their target populations as communities located in areas where such storms frequently occur.

A population at risk can be defined only in probabilistic terms. Women of childbearing age may be the population at risk for a program that provides birth control assistance, but a particular woman may or may not conceive a child within a given period of time. In this instance, specifying the population at risk simply in terms of age results unavoidably in overinclusion; many women who meet that definition will not need family planning services because they are not sexually active or are otherwise unlikely to become pregnant.

A target population may also be specified in terms of current need rather than risk, referred to as a **population in need**. Members of a population in need can be identified through direct assessments of their condition. For instance, there are reliable and valid literacy tests that can be used to identify functionally illiterate persons who constitute the population in need for adult literacy programs. For programs directed at alleviating poverty, the population in need may be defined as families whose annual incomes, adjusted for family size, are below a certain specified minimum. The fact that individuals are members of a population in need, however, does not necessarily mean that they want the program that serves that need. Desire for a service and willingness to participate in a program define the extent of the demand for a particular service irrespective of the attributed need. Community leaders and service providers, for instance, may define a need for residential facilities for the elderly when some significant number of

elderly persons do not want to use such facilities. Thus, need is not equivalent to demand.

Some needs assessments undertaken to estimate the extent of a problem are actually assessments of risk or assessments of demand rather than assessments of need according to the definitions just offered. For example, although only sexually active individuals are immediately appropriate for family planning services, the target population for most family planning programs is women at risk for unwanted pregnancies. It would be difficult and intrusive for a program to attempt to identify and designate only those who are sexually active as its target population. Similarly, whereas the in-need group for an evening literacy program may be all functionally illiterate adults, only those willing and able to participate can be considered the target population. The distinctions between populations at risk, in need, and at demand are therefore important for assessing the scope of a problem, estimating the size of the target population, and designing, implementing, and evaluating the program.

Incidence and Prevalence

Another useful distinction for describing the conditions a program aims to improve is the difference between incidence and prevalence. **Incidence** refers to the number of new instances of a particular problem that are identified or arise in a specified area or context during a specified period of time. **Prevalence** refers to the total number of existing cases in that area at a specified time. These concepts come from the field of public health, where they are sharply distinguished. To illustrate, the incidence of influenza during a particular month would be defined as the number of new cases reported that month. Its prevalence during that month would be the total number of people afflicted, regardless of when they were first stricken. In the health sector, programs generally are interested in incidence when dealing with disorders of short duration, such as upper respiratory infections and minor accidents. They are more interested in prevalence when dealing with problems that require long-term management and treatment, such as chronic conditions and long-term illnesses.

The concepts of incidence and prevalence also apply to social problems. In studying the impact of crime, for instance, a critical measure is the incidence of victimization: the number of new victims in a given jurisdiction over a defined period. Similarly, in programs aimed at lowering drunken-driving accidents, the incidence of such accidents may be the best measure of the need for intervention. But for chronic conditions such as low educational attainment, criminality, or poverty, prevalence is generally the more appropriate measure. In the case of poverty, for instance, prevalence may be defined as the number of poor individuals or families in a community at a given time, regardless of when they became poor.

Often, however, both prevalence and incidence are relevant for characterizing a target population. In dealing with unemployment, for instance, it is important to know its prevalence (the proportion of the population unemployed at a particular time). But the rate at which newly unemployed individuals enter that population is also of concern for programs that address unemployment.

Rates

In some circumstances it is useful to express incidence or prevalence as a **rate** within an area or population. Thus, crime victimization in a community during a given period might be described in terms of the percentage of persons victimized. Rates are especially appropriate for comparing problem conditions across areas or groups. For example, in describing crime victims, it is informative to have estimates by gender and age group. Although almost every age group is subject to some kind of crime victimization, married individuals and older persons are much less likely to be victims of serious crimes than their unmarried or younger counterparts. Such comparisons are meaningful when they are based on the proportions of the respective groups victimized but would be misleading if based on the number of such persons, because the groups are of quite different size. An alternative representation that allows consistent comparisons is a rate for a fixed number, for instance, the number of victimizations per thousand persons in the group or subgroup of interest.

[Exhibit 2-H](#) illustrates how prevalence rates can be used to characterize a target population in ways that identify the subgroups that are most likely to experience the problem at issue. For this example, crime victimization data from an annual national survey in the United States are broken down by gender, age, race/ethnicity, and marital status.

Exhibit 2-H Prevalence of Violent Crime by Demographic Characteristics of Victims

Victim Characteristics	Number of Persons Victimized	Prevalence Rate (%)
Total victimization in 2016	3,629,180	1.33
Gender		
Male	1,872,700	1.41
Female	1,756,490	1.26
Race/ethnicity		
White	2,318,090	1.34
Black	485,670	1.44
Hispanic	540,690	1.21
Other	284,730	1.34
Age (years)		
12–17	469,490	1.88
18–24	565,950	1.86
25–34	861,490	1.95
35–49	901,440	1.47
50–64	646,900	1.02
65 and older	183,910	0.38
Marital status		
Never married	1,822,230	1.90
Married	1,027,340	0.80
Widowed	96,920	0.64
Divorced	533,160	1.98
Separated	140,290	2.79
<i>Source: Morgan and Kena (2016).</i>		

Source: Morgan and Kena (2016).

Describing the Nature of Service Needs

As described above, a central function of needs assessment is to develop estimates of the extent and distribution of a given problem and the associated target population. However, it is also often important to develop descriptive information about the specific character of the need within that population. To be effective, a program must adapt its services to the local nature of the problem and the distinctive circumstances of the target population. This, in turn, requires information about the way in which the problem is experienced by those in that population, their perceptions and attributions about relevant services and programs, and the barriers and difficulties they encounter in attempting to access services.

A needs assessment might, for instance, probe why the problem exists and what other problems are linked with it. Investigation of low participation by high school students in Advanced Placement coursework may reveal that many schools do not offer such courses. Similarly, the incidence of depression among adolescents may be linked with high levels of cyberbullying. Consideration may also need to be given to cultural factors or perceptions and attributions that characterize a target population in ways that interact with their receptivity to program services. A needs assessment on poverty in rural populations, for instance, may highlight the sensitivities of the target population to accepting handouts and the strong value placed on self-sufficiency. Programs that are not consistent with these norms may be shunned to the detriment of the economic benefits they intend to facilitate.

Another important dimension of service needs may involve practical difficulties some members of the target population have in using services. This may result from transportation problems, limited service hours, lack of child care, or a host of similar obstacles. The difference between a program with an effective service delivery to persons in need and an ineffective one is often a matter of how much attention is paid to overcoming barriers such as these. Job training programs that provide child care to participants, nutrition programs that deliver meals to the homes of elderly persons, and community health clinics that are open during evening hours illustrate

approaches that have integrated awareness of access to service issues for their target populations into their program models.

Qualitative Methods for Describing Needs

Although many aspects of a needs assessment can be captured in quantitative data, qualitative research can be especially useful for obtaining detailed, textured knowledge of the specific needs in question. Such research can range from interviews of a few persons individually or in groups to elaborate and detailed ethnographic research. Carefully and sensitively conducted qualitative studies are particularly important for uncovering information with implications for how program services are configured. Qualitative studies of “no excuses” charter schools, for instance, will not only indicate how their disciplinary policies are experienced by students but will have implications for designing policies or programs that minimize disciplinary problems and enhance positive participation in the school culture. Or consider qualitative research on household energy consumption that might reveal how few householders know anything about the energy consumption characteristics of their appliances and thus have little capability to undertake effective strategies for reducing consumption. [Exhibit 2-I](#) provides an example of qualitative data on unmet needs for education and support among cancer survivors in American Indian and Alaska Native populations.

Exhibit 2-I Qualitative Data From a Needs Assessment on Cancer Education and Support in American Indian and Alaska Native Communities

Cancer is a leading cause of premature death for American Indian and Alaska Native populations. To inform public health efforts, a Web-based needs assessment survey focusing on unmet needs for cancer education and support was conducted by the Center for Clinical and Epidemiological Research at the University of Washington. Quantitative and qualitative data were collected from 76 community health workers and cancer survivors in northwestern United States. Content analysis of the qualitative responses to open-ended items asking about community needs for education and resources to assist cancer survivors identified three major themes:

- Resource needs
 - Need for psychosocial and logistical support for cancer survivors
 - Not enough money to pay for needed resources or services
- Barriers to receipt of health care services
 - Distance and lack of transportation
 - Fear and denial of illness
- Interest in information and communication

- Desire for face-to-face training and outreach
- Having print materials available to support training and outreach

The authors' overall conclusion was that their survey results highlighted the importance of culturally sensitive approaches to overcome barriers to cancer screening and education in American Indian and Alaska Native communities.

Source: Adapted from Harris, Van Dyke, Ton, Nass, and Buchwald (2016).

One useful technique for obtaining rich qualitative information about a social problem and its context is the **focus group**. Focus groups bring together selected persons for a discussion of a particular topic or theme facilitated by someone trained to elicit meaningful comments while minimizing conflict when disagreements arise. Appropriate participant groups generally include such stakeholders as knowledgeable community leaders, directors of service agencies, line personnel in those agencies who deal firsthand with clients, representatives of advocacy groups, and persons experiencing the social problem or service needs directly. With a careful selection and grouping of individuals, a modest number of focus groups can provide a wealth of descriptive information about the nature and nuances of a social problem and the service needs of those who experience it ([Exhibit 2-J](#) itemizes the steps for organizing a needs assessment focus group). A selection of other group-based techniques for eliciting needs assessment information can be found in Altschuld and Kumar (2010).

Exhibit 2-J Steps for Conducting a Needs Assessment Focus Group

The purpose of a focus group is to interview a group of individuals while promoting interaction among them on topics determined by the evaluator. Focus groups are useful for obtaining differentiated perspectives on the problem and needs, gaining clarity on those perspectives from the group interactions, assessing the extent to which views are commonly held or vary across individuals, getting fresh ideas from the participants, and building relationships and credibility for the findings. Steps for conducting a focus group include:

1. Determine that focus group interviews are appropriate for collecting the data needed for the needs assessment. Considerations include whether information that is needed can be collected from individuals in a group setting and if the validity of the information may be improved through group interactions.
2. Select individuals for the focus group interview. Considerations include identifying types of individuals who have firsthand information on the problem or existing attempts to ameliorate the problem and selecting a relatively homogeneous group for each focus group while achieving diversity through conducting multiple focus groups.

3. Attend to the logistical details and arrangements for making the focus group successful. Considerations include inviting participants sufficiently in advance, selecting a convenient time and place, providing comfortable seating that encourages interactions, and identifying a moderator prepared to lead the group and another individual to take notes and assist the moderator.
4. Prepare questions for the focus group. Considerations include phrasing questions about the problem, its causes, consequences, barriers to ameliorating it, and perspectives on current attempts to reduce it that can be answered in an open-ended manner by participants.
5. Conduct the focus group. Considerations include familiarity of the moderator with the topics, specific questions and moving through the questions in the allotted time, probing for additional depth and clarity, keeping all participants engaged, summarizing what has been heard to ensure clarity, and actively assessing the extent of agreement among the responses.
6. Analyze and report the findings. Considerations include identifying the main ideas within and across the focus groups, determining the themes that arose in the responses, and organizing communication of the themes to stakeholders.

Source: Adapted from Altschuld (2010).

Any use of key informants in needs assessment must involve a careful selection of the persons or groups whose perceptions will be taken into account. A useful way to identify such informants is **snowball sampling**, in which an initial set of informants is located and asked to identify other informants whom they believe to be knowledgeable about the matter at issue. Those informants, in turn, are also asked to identify other appropriate informants. When this process no longer produces relevant new names it is likely that most of those who would qualify as key informants have been identified. In many cases incentives for participation, such as a nominal payment, are provided to those who agree to participate and for everyone they recruit who agrees to participate.

However, asking informants to identify other individuals may invade the privacy of those others or put them at risk for unwanted disclosure of their circumstances. A modification of this procedure is for the evaluator to provide the initial informants with information on how to contact the evaluation team and ask them to recruit other key informants. These other informants are unknown to the evaluation team until they initiate contact and express willingness to participate. They are then interviewed and asked, in turn, to recruit still others using the same procedure by which they were recruited.

An especially useful group of informants that should not be overlooked in needs assessment consists of a program's current clientele or, in the case of a new program, representatives of its potential clientele. This group, of course, is especially knowledgeable about the characteristics of the problem and the associated needs as they are experienced by those whose lives are affected by the problem. Although they are not in the best position to report on how widespread the problem is, they are key witnesses with regard to how seriously the problem affects individuals and what dimensions of it are most pressing. Care must be taken to protect the privacy of key informants who are clients or prospective clients for the program at issue. Identification of persons who are members of the target population implies that they experience the problem the program addresses, which may be a sensitive matter. For instance, identification of users of illegal opioids may place those individuals at risk for unwanted attention from authorities. Similarly, mental health patients may not want that status revealed to employers.

Because of the distinctive advantages of qualitative and quantitative approaches, a useful and frequently used strategy is to conduct needs assessment in two stages. The initial, exploratory stage uses qualitative techniques to obtain rich information on the nature of the problem. The second stage builds on this information to design a quantitative assessment that provides reliable estimates of the extent and distribution of the problem as well as more exact information about the experience of the target population with the different aspects of the problem identified in the qualitative data.

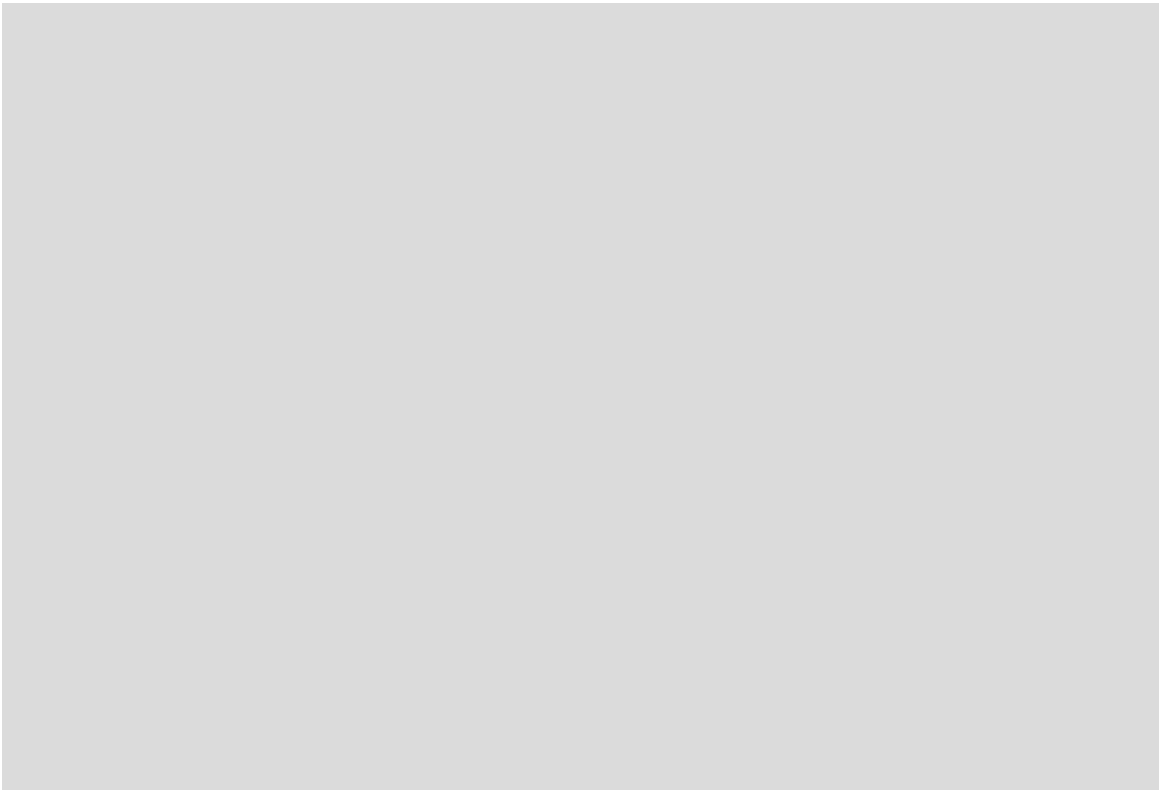
Summary

- Needs assessment answers questions about the need for a program and the social conditions it is intended to address, or whether a new program is needed. More generally, needs assessment may be used to identify, compare, and prioritize needs within and across program areas.
- Adequate diagnosis of social problems, identification of the target population for intervention, and description of the characteristics of the target population that have implications for appropriate services and service delivery are prerequisites for the design and operation of effective programs.
- Social problems are not objective phenomena; rather, they are social constructs that emerge from social and political agenda-setting processes. Evaluators can play a useful role in assisting policymakers and program managers to refine the definitions of the social problems in ways that allow intervention to be appropriate and effective.

- To specify the size, distribution, and characteristics of a problem, evaluators may gather and analyze data from existing sources, such as government-sponsored surveys, censuses, and social indicators. Because some or all of the information needed often cannot be obtained from such sources, evaluators frequently collect their own needs assessment data. Useful sources of data for that purpose include agency records, sample surveys, key informant interviews, and focus groups.
- Forecasts of future needs are often relevant to needs assessment but generally involve considerable uncertainty and are typically technical endeavors conducted by specialists. In using forecasts, evaluators must take care to assess the assumptions and data on which the forecasts are based.
- The target population for a program may be individuals, groups, geographic areas, or physical units, and they may be defined as direct or indirect objects of an intervention. Specification of the membership of a target population should establish appropriate boundaries that are feasible to apply and that allow interventions to correctly identify and serve that population.
- Useful concepts for defining target populations include *population at risk*, *population in need*, *population at demand*, *incidence and prevalence*, and *rates*.
- For purposes of program planning or evaluation, it is important to have detailed, contextualized information about the local nature of a social problem and the distinctive circumstances of those in need. Such information is often best obtained through qualitative methods such as ethnographic studies, key informant interviews, or focus groups with representatives of various stakeholders and program participants.

Key Concepts

Focus group 53
Incidence 50
Key informants 45
Needs assessment 32
Population at risk 49
Population in need 49
Prevalence 50
Probability sample 38
Rate 50
Sample survey 41
Sampling frame 38
Snowball sampling 55
Social indicator 38
Target population 32
Targeted program 47
Universal program 47



Critical Thinking/Discussion Questions

1. This chapter outlines six probability sampling designs. Explain each sampling design and state when each is appropriate to be used.
2. Three types of data sources from which evaluators can obtain pertinent needs assessment data are described in this chapter. Discuss each one and explain when it would be applicable to use in a needs assessment.
3. Explain how a target population is identified in an evaluation. Choose three important considerations in identifying a target population and discuss how researchers must deal with these challenges.

Application Exercises

1. [Exhibit 2-A](#), “The Three Phases of Needs Assessments,” outlines the needs assessment process. Locate a published needs assessment and identify how the researchers addressed the components included in each of the three phases.
 - Phase 1: Preassessment
 - Phase 2: Assessment
 - Phase 3: Postassessment
2. Identify a social problem to research, then find a nationally representative survey to use as your data source. List the key social indicators included in the data set that you will use in your analysis. How are these social indicators measured in the data set you have chosen? What social indicators would you like to include but cannot as they are not measured in the data set?

Chapter 3 Assessing Program Theory and Design

[Evaluability Assessment](#)

[Describing Program Theory](#)

[Program Impact Theory](#)

[Service Utilization Plan](#)

[Organizational Plan](#)

[Eliciting Program Theory](#)

[Defining the Boundaries of the Program](#)

[Explicating the Program Theory](#)

[Program Goals and Objectives](#)

[Program Functions, Components, and Activities](#)

[The Logic or Sequence Linking Program Functions,](#)

[Activities, and Components](#)

[Corroborating the Description of the Program Theory](#)

[Assessing Program Theory](#)

[Assessment in Relation to Social Needs](#)

[Assessment of Logic and Plausibility](#)

[Assessment Through Comparison With Research and Practice](#)

[Assessment via Preliminary Observation](#)

[Possible Outcomes of Program Theory Assessment](#)

[Summary](#)

[Key Concepts](#)

The social problems addressed by programs are often so complex and difficult that bringing about even small improvements may pose formidable challenges. A program's theory is the conception of what must be done to bring about the intended changes. As such, it is the foundation on which every program rests.

A program's theory can be a sound one, in which case it represents the understanding necessary for the program to attain the desired results, or it can be a poor one that would not produce the intended effects even if implemented well. One aspect of evaluating a program, therefore, is to assess how good the program theory is—in particular, how well it is formulated and whether it presents a plausible and feasible plan for bringing about the intended improvements. For program theory to be assessed, however, it must first be

expressed clearly and completely enough to stand for review. Accordingly, this chapter describes how evaluators can describe program theory and then assess how sound it is.

Mario Cuomo, former governor of New York, once described his mother's rules for success as (a) figure out what you want to do and (b) do it. These are pretty much the same rules social programs must follow if they are to be effective. Given an identified need, program decision makers must (a) conceptualize a program capable of alleviating that need and (b) implement it. In this chapter, we review the concepts and procedures an evaluator can apply to the task of assessing the quality of the program conceptualization, which is often referred to as the program theory. In [Chapter 4](#), we describe how the evaluator can assess the program's implementation.

Whether it is expressed in a detailed program plan and rationale or is only implicit in the program's structure and activities, the program theory explains why the program does what it does and provides the rationale for expecting that doing so will achieve the desired results. When examining a program's theory, evaluators may find that it is not very convincing. There are many poorly designed social programs with faults that reflect deficiencies in their underlying conceptions of how the desired social benefits can be attained. This happens in large part because insufficient attention is given during the planning of new programs to carefully conceptualizing their objectives and how those objectives are supposed to be achieved. Sometimes the political context does not permit extensive planning, but even when that is not the case, conventional practices for designing programs pay little attention to the underlying theory. The human service professions operate with repertoires of established services and types of intervention associated with their respective specialty areas. As a result, program design is often a matter of configuring a variation of familiar services into a package that seems appropriate for a social problem without a close analysis of the match between those services and the specific nature of the problem.

For example, many social problems involve risky behavior, such as alcohol or drug abuse, criminal behavior, early sexual activity, or teen pregnancy, that frequently are addressed by programs that provide the target populations with some mix of counseling and educational services. This approach is based on an assumption that is rarely made explicit during the

planning of the program, namely, that people will change their problem behaviors if given information and interpersonal support for doing so. Although this assumption may seem reasonable, experience and research provide ample evidence that such behaviors are resistant to change even when participants know they should change and receive strong encouragement to do so. Thus, the theory that education and supportive counseling by themselves will reduce risky behavior may not be a sound basis for program design.

A program's rationale and conceptualization, therefore, are just as subject to critical scrutiny within an evaluation as any other important aspect of the program. If the program's goals and objectives do not relate in a reasonable way to the social conditions the program is intended to improve, or the assumptions and expectations embodied in the program's design do not represent a credible approach to bringing about that improvement, there is little prospect that the program will be effective.

The first step in assessing program theory is to articulate it, that is, to produce an explicit description of the conceptions, assumptions, and expectations that constitute the rationale for the way the program is structured and operated. Only rarely can key program stakeholders immediately provide the evaluator with a full statement of its underlying theory. Although the program theory is always implicit in the program's structure and operations, a detailed account is seldom written down in program documents. Moreover, even when some write-up of program theory is available, it is often in material prepared for funding proposals or public relations purposes and may not correspond well with actual program practice.

Assessment of program theory, therefore, almost always requires that the evaluator first synthesize and articulate the theory in a form amenable to analysis. Accordingly, the discussion in this chapter is organized around two themes: (a) how the evaluator can explicate and express program theory in a form that will be representative of key stakeholders' actual understanding of the program and workable for purposes of evaluation and (b) how the evaluator can assess the quality of the program theory that has been thus articulated. We begin with a brief description of a set of

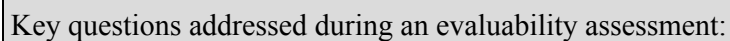
evaluative activities known collectively as evaluability assessment that are frequently implemented to develop the program theory and determine the feasibility of an evaluation.

Evaluability Assessment

As the evaluation of social programs became more commonplace, many evaluators found it difficult to design informative evaluations of some of the programs they were charged with assessing. The barriers to conducting useful evaluations they identified included stakeholder disagreement about the goals and objectives of the program or, when there was agreement, program activities and resources that were not sufficient to have a reasonable chance to accomplish the program aims. In other cases, key program decision makers were not open to making program changes on the basis of evaluation findings. This led to the view that a qualitative assessment of whether minimal preconditions for evaluation were met should precede most evaluation efforts. Joseph Wholey (1987, 2015), who articulated this approach, termed the process **evaluability assessment**, and it has become a widely used tool for systematic evaluation planning. The aims and process for conducting an evaluability assessment are described in [Exhibit 3-A](#).

Exhibit 3-A Rationale for Evaluability Assessment

Evaluability assessments are undertaken to ensure that a program is ready to be evaluated before committing to do so. Leviton, Khan, Rog, Dawkins, and Cotton (2010) diagrammed the process of evaluability assessments in a way that highlights several important questions about the preconditions necessary to conduct an evaluation, using arrows to identify parts of the assessment process that may require iterating between steps before moving forward.



1. Is there agreement on goals and objectives for the program? If stakeholders disagree on the program's goals and objectives, the program is not ready to be evaluated.
2. Has the logic underlying the program or practice been described in sufficient detail to explain how the program is expected to achieve its goals and objectives? If not, the evaluator will need to create a logic model or program theory on which stakeholders agree, or the program logic will need to be further developed.
3. Is it plausible that the program can accomplish its goals and objectives? Staff may be able to describe program logic, but goals and objectives may not be realistic given the resources available or the activities being undertaken. At this point, an evaluability assessment can indicate the need for further program development or, possibly, a formative evaluation.
4. Do key stakeholders agree about performance criteria or how to measure program effectiveness? Stakeholders need to agree on the criteria by which a program's effectiveness will be judged before an influential evaluation measuring those criteria can be conducted.
5. Can the program or evaluation sponsor afford the cost of an evaluation?
6. Do key stakeholders agree on the relevance of a program evaluation and indicate willingness to make changes to the program on the basis of the evaluation? If the stakeholders are not open to making changes, the utility of an evaluation is doubtful.

In addition to addressing these questions, evaluability assessments often determine if the data needed to carry out an evaluation on the basis of the performance criteria are obtainable. The arrow in the figure that is directed back to "Create/revise logic model or theory of change" indicates that the evaluability assessment may identify aspects that may require revision.

Source: Leviton, Khan, Rog, Dawkins, and Cotton (2010).

Evaluability assessment involves three primary activities: (a) description of the *program model*, with particular attention to defining the program goals and objectives; (b) assessment of how well defined and evaluable that model is; and (c) identification of stakeholder interest in evaluation and the likely use of the findings. Evaluators conducting evaluability assessments operate much like ethnographers in that they seek to describe and understand the program through interviews and observations that will reveal its social reality as viewed by program personnel and other significant stakeholders. The evaluators begin with the conception of the program presented in documents and official information, but then try to see the program through the eyes of those closest to it. The intent is to end up with a description of the program as it exists and an understanding of the program issues that really matter to the parties involved. Although this process involves considerable judgment and discretion on the part of the evaluator, various practitioners have attempted to codify its procedures so

that evaluability assessments will be reproducible by other evaluators (see Davies, 2013; Thurston & Potvin, 2003; Wholey, 2015).

A common outcome of evaluability assessments is that program managers and sponsors recognize the need to modify their programs. The evaluability assessment may reveal faults in a program's delivery system, that the program's target population is not well defined, or that the intervention itself needs to be redesigned. Or there may be few program objectives stakeholders agree on or no feasible performance indicators for the objectives. In such cases, the evaluability assessment has uncovered problems with the program's design that program managers must correct before any meaningful performance evaluation can be undertaken.

The aim of evaluability assessment is to create a favorable climate and an agreed-on understanding of the nature and objectives of the program that will facilitate the design of an evaluation. As such, it can be integral to the approach the evaluator uses to tailor an evaluation and formulate evaluation questions (see [Chapter 1](#)). [Exhibit 3-B](#) presents an example of an evaluability assessment that illustrates a very systematic procedure due to the scope of the assessment: examining 40 developmental cooperation interventions to gain awareness of the obstacles to evaluation in this field.

Evaluability assessment requires program stakeholders to articulate the program's design and logic (the program model); however, it can also be carried out for the purposes of describing and assessing program theory (Wholey, 1987). Indeed, the evaluability assessment approach represents the most fully developed set of concepts and procedures available in the evaluation literature for describing and assessing a program's design, including what it is supposed to be doing and why. We turn now to a more detailed discussion of procedures for identifying and evaluating program theory.

Exhibit 3-B Evaluability Assessment of Belgian Development Cooperation

The evaluators conducted an evaluability assessment of 40 development interventions financed through the Belgium Development Agency, Belgian nongovernmental organizations (NGOs), or other agencies in the country. To be systematic across the interventions, they developed a framework consisting of three overarching dimensions: (a) analysis of the intervention design, including the underlying theory of change; (b)

practice regarding intervention implementation, intervention management, and context, including availability of information regarding the implementation and results of the intervention as well as the activity monitoring system in practice; and (c) the evaluation context, focusing on its conduciveness to evaluation. Under these three dimensions, a total of 62 indicators of the evaluability of the interventions were rated on a common scale. In addition, evaluability for different types of evaluation were assessed, for instance, impact evaluation, assessment of cost and efficiency, and the sustainability of benefits after the development assistance has been completed.

During the conduct of the evaluability assessment, the evaluators collected secondary and primary data. Secondary data included

- intervention proposal,
- baseline report,
- progress reports (e.g., midterm reports, yearly reports, and end-term reports),
- prior studies and evaluations, and
- monitoring and evaluation policy documents of the organizations involved.

Primary data collection included

- four focus group discussions at the headquarters of the organizations financing the development interventions in Brussels, the Belgian Development Agency, Directorate General of Development Cooperation, and NGOs;
- site visits to the four countries from which the evaluators drew their study sample (Republic of the Congo, Benin, Rwanda, and Belgium); and
- interviews with 15 to 25 individuals at each program site.

The evaluators found that the intervention logic and the theory of change were rated highly for evaluation of efficiency and achievement of the implementation objectives, but lower for impact evaluation. With respect to data availability, the assessment found that available data were appropriate for evaluating the achievement of the interventions' objectives and costs but not for evaluating impact or sustainability.

Overall, the evaluators raised concerns that elements that would support impact evaluation, such as baseline information on a group that could be used to compare the outcomes of the intervention, were not developed when the interventions began, making credible impact evaluation less feasible. A major contribution of the study was making the criteria to be used to assess evaluability explicit and transparent and developing rubrics that facilitated reliable scoring.

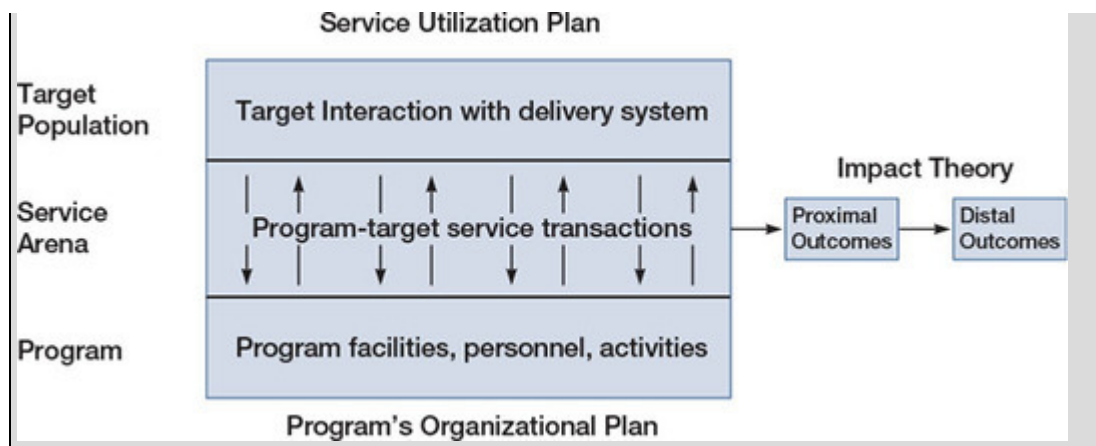
Source: Adapted from Holvoet et al. (2018).

Describing Program Theory

Evaluators have long recognized the importance of program theory as a basis for formulating and prioritizing evaluation questions, designing evaluation research, and interpreting evaluation findings (Bickman, 1987; Chen & Rossi, 1980; Weiss, 1972, 1997; Wholey, 1979) and the developments continue apace (Christie & Alkin, 2003; Donaldson, 2007). However, program theory has been described and used under various names, for example, logic model, program model, outcome line, cause map, and action theory. There is no consensus about how best to describe a program's theory, so we will present a scheme we have found useful in our own evaluation activities.

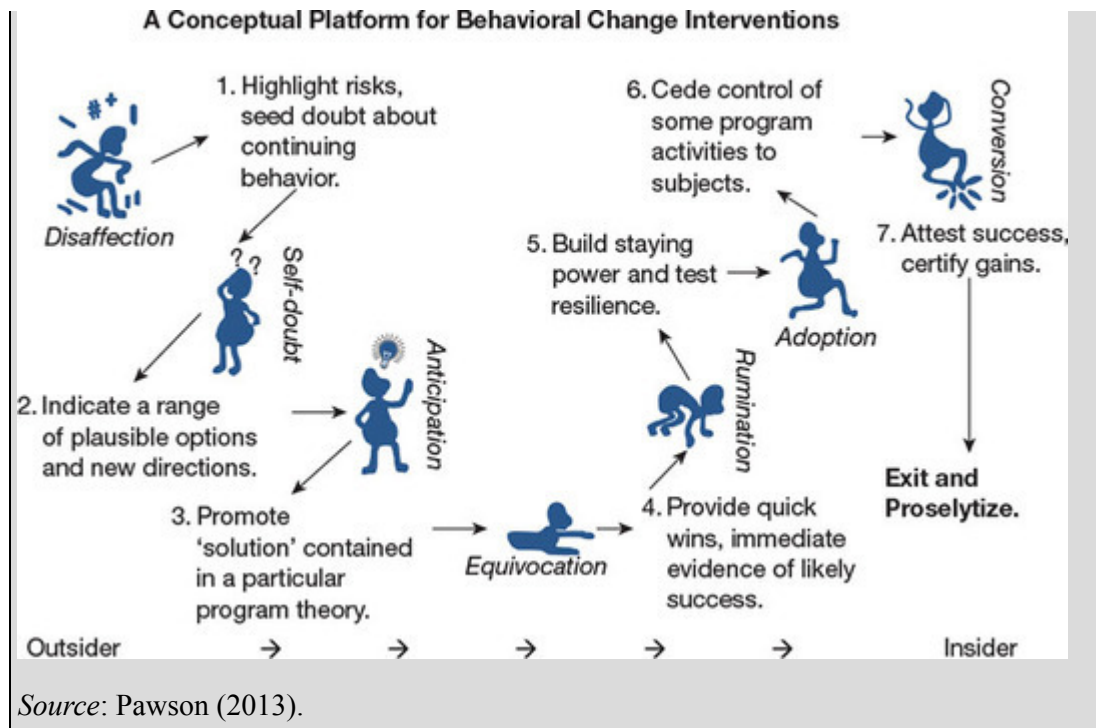
For this purpose, we depict a social program as centering on the transactions that take place between a program's operations and the target population it serves ([Exhibit 3-C](#)). These transactions might involve counseling sessions for women with eating disorders in therapists' offices, recreational activities for high-risk youths at a community center, educational presentations to local citizens' groups, nutrition posters in a clinic, informational pamphlets about empowerment zones and tax law mailed to potential investors, delivery of meals to the front doors of elderly persons, or any such point-of-service contact. On one side of this program–target population transaction, we have the program as an organizational entity with its various facilities, personnel, resources, activities, and so forth. On the other side, we have the target participants in their life spaces with their various circumstances and experiences in relation to the service delivery system of the program.

Exhibit 3-C Overview of Program Theory



This simple scheme highlights three interrelated components of program theory: the program impact theory, the service utilization plan, and the program's **impact theory**, also referred to as a theory of change, consists of assumptions about the change process actuated by the program and the outcomes that are expected to be effected as a result. That change process is operationalized by the program–target population transactions, for they constitute the means by which the program expects to bring about the intended effects. The impact theory may be as simple as presuming that exposure to information about the negative effects of drug abuse will motivate high school students to abstain or as complex as the ways in which an eighth grade science curriculum will lead to deeper understanding of natural phenomena. It may be as informal as the commonsense presumption that providing hot meals to elderly persons improves their nutrition or as formal as classical conditioning theory adapted to treating phobias. Whatever its nature, however, an impact theory constitutes the essence of a social program. If the assumptions embodied in that theory about how desired changes are brought about by program action are faulty, or if they are valid but not well operationalized by the program, the intended social benefits will not be achieved.

Exhibit 3-D Program Impact Theory: Realizing Positive Behavioral Change



When evaluating a program impact theory, evaluators must assess whether it can realistically produce the expected changes required to realize the program goals and objectives. In most cases, programs must change individuals' behaviors in order to work effectively. In education, professional development workshops are expected to change the way teachers instruct their students in order for students to learn more. Crime deterrence programs are expected to reduce the criminal behaviors of individuals who commit crimes. In [Exhibit 3-D](#), a social science theory of behavioral change as an individual moves from outsider to insider status is laid out in seven stages (Pawson, 2013). *Outsider* refers to someone whose behaviors are outside that desired to fulfill the program goals, and *insider* refers to someone whose behaviors help realize the program's goals. In this theory, the individual who is currently acting contrary to the program goals begins to question those behaviors, then to anticipate behaving in a way that is being promoted by the program. Engaging in the behavior that produces positive outcomes in the form of quick wins then promotes adoption of the behavior and conversion to an insider who behaves in a manner consistent with achieving the program's goals. Pawson describes this as a basic model of behavioral change that can be adapted and applied to numerous types of programs, often after close inspection of the behaviors of individuals receiving services as they interact with the program personnel. Other social

science theories, such as the theory of planned behavior (Ajzen & Fishbein, 1980), have also been adapted for programs that target behavioral change.

To instigate the change process posited in the program's impact theory, the intended services must first be provided to the target population. The program's **service utilization plan** includes the program's assumptions and expectations about how to reach the target population, provide and sequence service contacts, and conclude the relationship when services are no longer needed or appropriate. For a program to increase awareness of AIDS risk, for instance, the service utilization plan may be simply that appropriate persons will read informative posters if they are put up in subway cars. A multifaceted AIDS prevention program, on the other hand, may be organized on the assumption that high-risk drug abusers who are referred by outreach workers will go to nearby street-front clinics, where they will receive appropriate testing and information.

The program, of course, must be organized in such a way that it can actually provide the intended services. The third component of program theory, therefore, relates to program resources, personnel, administration, and general organization. We call this component the program's **organizational plan**. The organizational plan can generally be represented as a set of propositions: If the program has such and such resources, facilities, personnel, and so on, if it is organized and administered in such and such a manner, and if it engages in such and such activities and functions, then a viable organization will result that can operate the intended service delivery system. Elements of programs' organizational plans include, for example, assumptions that case managers should have master's degrees in social work and at least 5 years' experience, that at least 20 case managers should be employed, that the agency should have an advisory board that represents local business owners, that an administrative coordinator should be assigned to each site, and that working relations should be maintained with the Department of Public Health.

Adequate resources and effective organization, in this scheme, are the factors that make it possible to develop and maintain a service delivery system that enables use of the services by the target population. A program's organization and the service delivery system that organization

supports are the parts of the program most directly under the control of program administrators and staff. These two aspects together are often referred to as *program process*, and the assumptions and expectations on which that process is based may be called the program **process theory** or the theory of action.

With this overview, we turn now to a more detailed discussion of each of the components of program theory, with particular attention to how the evaluator can describe them in a manner that permits analysis and assessment.

Program Impact Theory

Program impact theory is causal theory. It describes a cause-and-effect sequence in which certain program activities are the instigating causes and certain social benefits are the effects they eventually produce. These theories can be rooted in social science, as in the behavioral change theories above, or more pragmatic ways of describing the interrelationships between programmatic actions and changes that lead to the desired program outcomes. Evaluators typically represent program impact theory in the form of a causal diagram showing the cause-and-effect linkages presumed to connect a program's activities with the expected outcomes (Chen, 1990; Lipsey, 1993; Martin & Kettner, 1996). Because programs rarely exercise direct control over the social conditions they are expected to improve, they must generally work indirectly by changing some critical but manageable aspect of the situation, which, in turn, is expected to lead to more far reaching improvements.

The simplest program impact theory is the basic “two-step,” in which services affect some intermediate condition that, in turn, improves the social conditions of concern (Lipsey & Pollard, 1989). For instance, a program cannot make it impossible for people to abuse alcohol, but it can attempt to change their attitudes and motivation toward alcohol in ways that provide them with the support necessary to avoid abuse. More complex program theories may have more steps along the path between program and social benefit, as in the seven-stage behavioral change model (Pawson, 2013), and, perhaps, involve more than one distinct path.

The distinctive features of any representation of program impact theory are that each element is a cause-effect link in a chain of events that begins with program actions and ends with change in the outcomes the program intends to improve (see [Exhibit 3-E](#)). The events following directly from the instigating program activities are the most direct outcomes, often called proximal or immediate outcomes (e.g., dietary knowledge and awareness in the first example in [Exhibit 3-E](#)). Events further down the chain constitute the more distal or ultimate outcomes (e.g., healthier diet in the first example in [Exhibit 3-E](#)). Program impact theory highlights the dependence of the

more distal, and generally more important, outcomes on successful attainment of the more proximal ones.

Service Utilization Plan

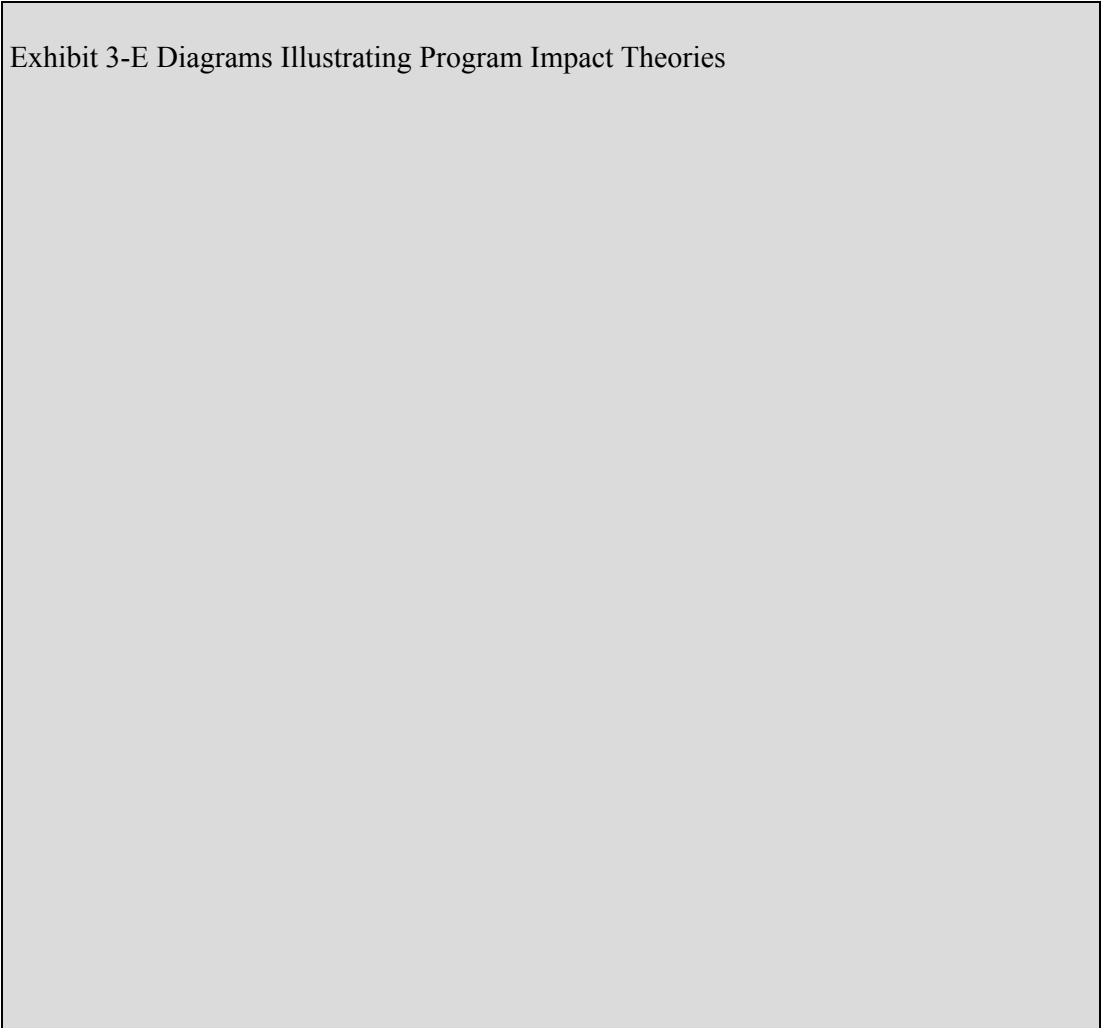
An explicit service utilization plan pulls into focus the critical assumptions about how and why the intended recipients of service will actually become engaged with the program and follow through to the point of receiving sufficient services to initiate the change process represented in the program impact theory. It describes the program–target population transactions from the perspective of the program participants and their life spaces as they might encounter the program.

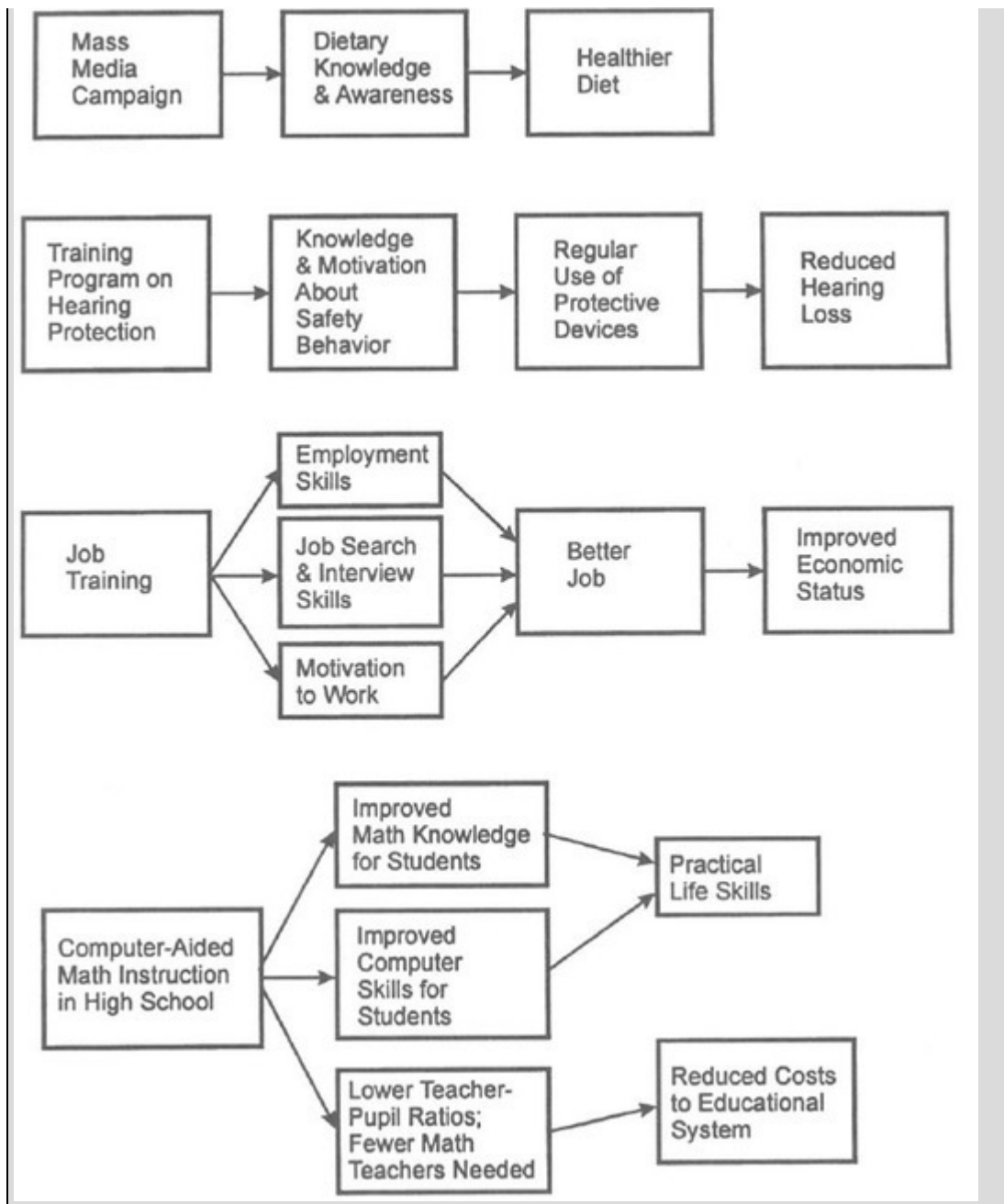
A program's service utilization plan can be usefully depicted in a flowchart that tracks the various paths program participants can follow from some appropriate point prior to first contact with the program through a point at which there is no longer any contact. [Exhibit 3-F](#) shows an example of a simple service utilization flowchart for a hypothetical aftercare program for released psychiatric patients. One characteristic of such charts is that they identify the possible situations in which the program targets are not engaged with the program as intended. In [Exhibit 3-F](#), for example, we see that formerly hospitalized psychiatric patients may not receive the planned visit from a social worker or referrals to community agencies and, as a consequence, may receive no service at all.

Organizational Plan

The program's organizational plan is articulated from the perspective of program management. The plan encompasses both the functions and activities the program is expected to perform and the human, financial, and physical resources required for that performance. Central to this scheme are the program services: those specific activities that constitute the program's role in the program–target population transactions expected to lead to social benefits. However, the organizational plan also must include those functions that provide essential preconditions and ongoing support for the organization's ability to provide its primary services, for instance, fund-raising, personnel management, facilities acquisition and maintenance, political liaison, and the like.

Exhibit 3-E Diagrams Illustrating Program Impact Theories

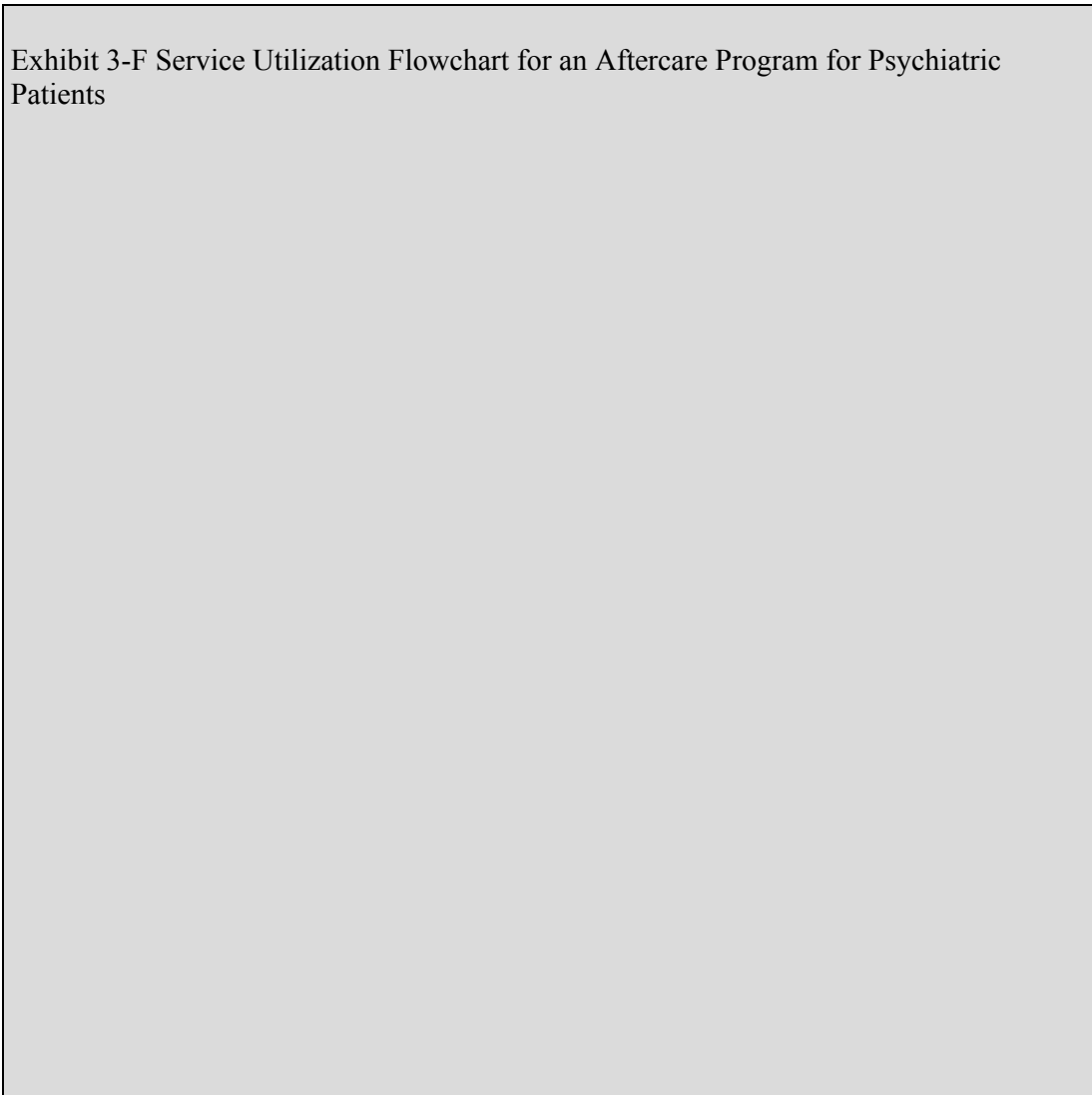




There are many ways to depict a program's organizational plan. If we center it on the program–target population transactions, the first element will be a description of the program's objectives for the services it will provide: what those services are, how much is to be provided, to whom, on what schedule, and so on. The next element might then describe the resources and functions necessary to engage in those service activities, for instance, sufficient personnel with appropriate credentials and skills, proper facilities and equipment, funding, supervision, clerical support, and so forth.

As with the other portions of program theory, it is often useful to describe a program's organizational plan with a diagram. [Exhibit 3-G](#) presents an example that depicts the major organizational components of the aftercare program for psychiatric patients whose service utilization scheme is shown in [Exhibit 3-F](#). A common way of representing the organizational plan of a program is in terms of inputs (resources and constraints applicable to the program) and activities (services the program is expected to provide). In a full logic model of the program, receipt of services (service utilization) is represented as program outputs, which, in turn, are related to the desired outcomes. [Exhibit 3-H](#) shows an appropriately detailed logic model for improving children's healthy eating habits and physical activity that addresses both the impact theory and organization and service elements of the program.

Exhibit 3-F Service Utilization Flowchart for an Aftercare Program for Psychiatric Patients



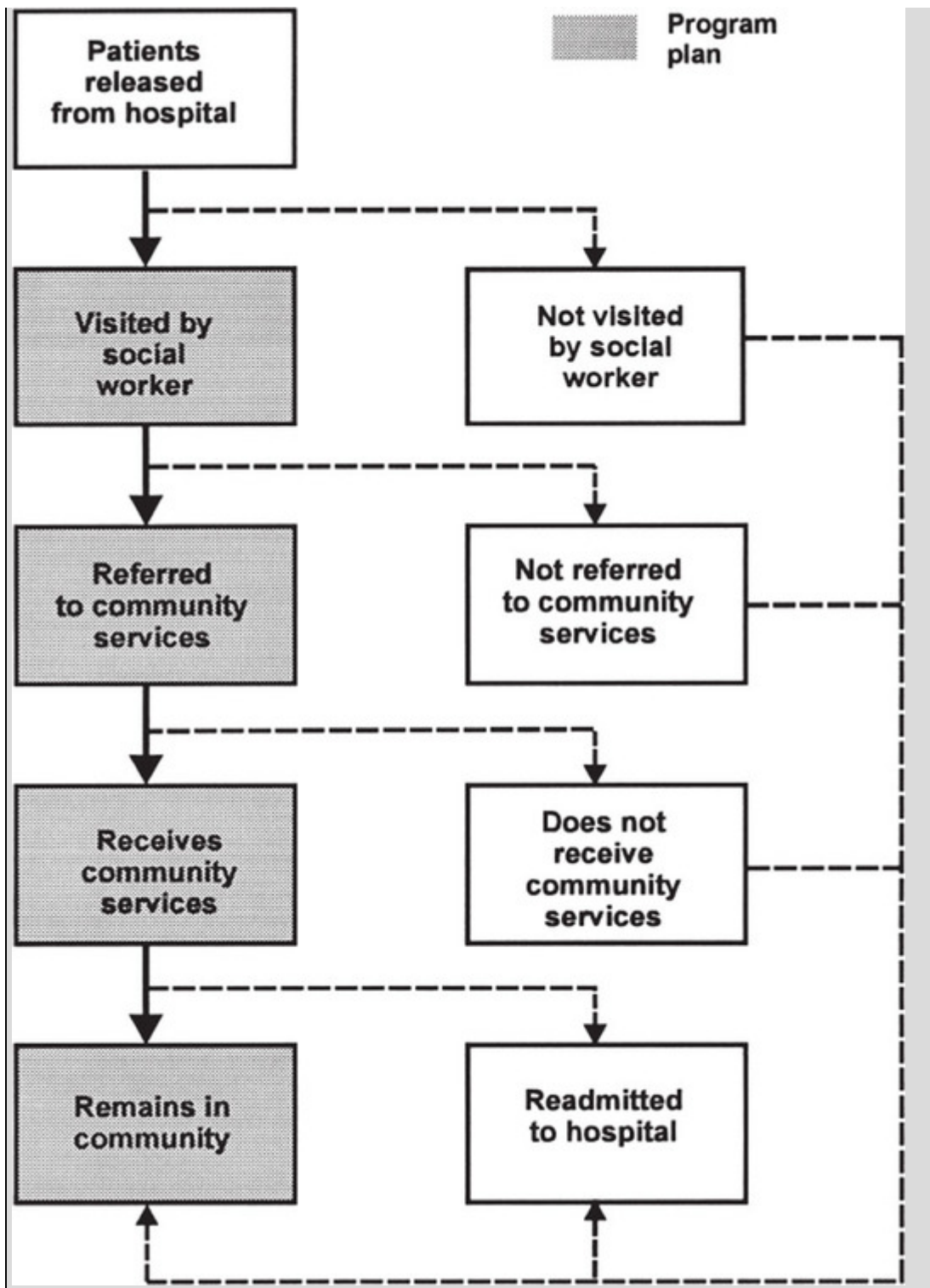
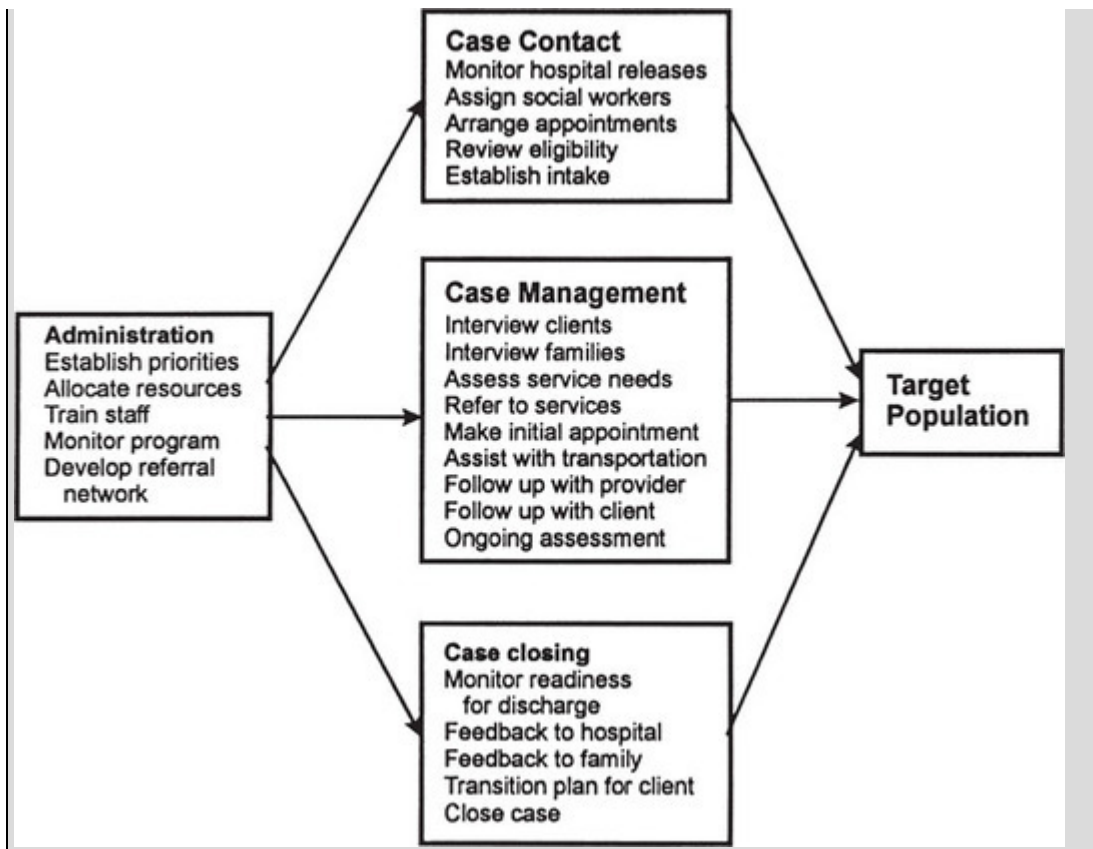


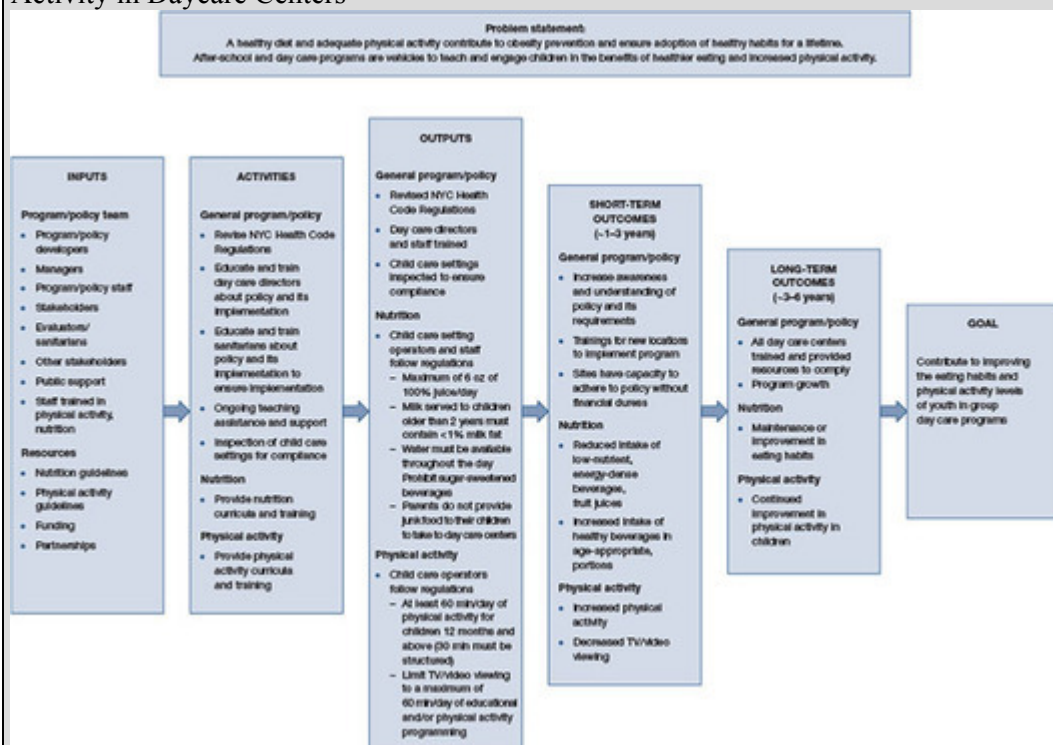
Exhibit 3-G Organizational Schematic for an Aftercare Program for Psychiatric Patients



Eliciting Program Theory

Carol Weiss, one of the pioneers of program evaluation, made numerous contributions to evaluators' understanding of program theory and how to elicit a program theory (see [Exhibit 3-I](#) for some of her contributions to program theory). When a program's theory is spelled out in program documents and well understood by staff and stakeholders, the program is said to be based on an **articulated program theory** (Weiss, 1997). This is most likely to occur when the original design of the program is drawn from social science theory. For instance, a school-based drug use prevention program that features role-playing of refusal behavior in peer groups may be derived from social learning theory and its implications for peer influences on adolescent behavior.

Exhibit 3-H A Logic Model for a Program That Promotes Healthy Eating and Physical Activity in Daycare Centers



Source: Leviton et al. (2010).

Exhibit 3-I Carol Weiss: Evaluation Pioneer and Contributor to the Concept of Program Theory

Carol Weiss



Theory-based evaluation is demonstrating its capacity to help readers understand how and why a program works or fails to work. Knowing only outcomes, even if we know them with irreproachable validity, does not tell us enough to inform program improvement or policy revision. Evaluation needs to get inside the black box and do so systematically. . . . Probably the central need is for better program theories. Evaluators are currently making do with the assumptions that they are able to elicit from program planners and practitioners or with the logical reasoning that they bring to the table. . . . Evaluators need to look to the social sciences, including social psychology, economics, and organization studies, for clues to more valid formulations. . . . Better theories are even more essential for program designers, so that social interventions have a greater likelihood of achieving the kind of society we hope for in the twenty-first century. (Weiss, 1997)

Carol Weiss, who passed away in 2013, was the Beatrice Whiting Professor Emeritus in the Harvard University Graduate School of Education, where she had worked since 1978. In one of her foundational contributions to the theory-driven approach to evaluation, Weiss made explicit the differences between theories surrounding implementation and theories that explore underlying mechanisms necessary to ensure programs work as intended. She referred to a combination of both theories as “theories of change.” The identification and measurement of change mechanisms are a key feature of Weiss’s work on program theory, that is, not only enumerating and measuring the variables identified in the causal chain but also measuring mediating variables that explain how the causal process works (Weiss, 1997).

In addition to theory-driven evaluation, Weiss contributed extensively to understanding the influence and use of evaluation. Using the term *enlightenment*, Weiss posited that evaluation, and social science research more broadly, provides us with ways of understanding social programs, the problems they address, and the conditions they are expected to ameliorate (Weiss, 1979).

The contributions of Carol Weiss remain hugely relevant and influential in the field of evaluation research. The link between enlightenment and program theory highlights the conceptual dimension of programs and the associated implications for program evaluation and its role in guiding policy and practice.

Sources: Weiss (1972, 1979, 1997).

When the underlying assumptions about how program services and practices are presumed to accomplish their purposes have not been fully articulated and recorded, the program has an **implicit program theory** or, as Weiss (1997) put it, a tacit theory. This might be the case for a counseling program to assist couples with marital difficulties. Although it may be reasonable to assume that discussing marital problems with a trained professional would be helpful, the way in which that translates into improvements in the marital relationship is not described by an explicit theory, nor would different counselors necessarily agree about the process.

When a program's theory is implicit rather than articulated, the evaluator must extract and describe it before it can be analyzed and assessed. The evaluator's objective is to depict the "program as intended," that is, the actual expectations held by decision makers about what the program is supposed to do and what results are expected to follow. With this in mind, we now consider the concepts and procedures an evaluator can use to extract and articulate program theory as a prerequisite for assessing it.

Defining the Boundaries of the Program

A crucial early step in articulating program theory is to define the boundaries of the program at issue (Smith, 1989). A human service agency may have many programs and provide multiple services; a regional program may have many agencies and sites. There is usually no one correct definition of a program, and the boundaries the evaluator applies will depend, in large part, on the scope of the evaluation sponsor's concerns and the program domains to which they apply.

One way to define the boundaries of a program is to work from the perspective of the decision makers who are expected to act on the findings of the evaluation. The evaluator's definition of the program should at a minimum represent the relevant jurisdiction of those decision makers and the organizational structures and activities about which decisions are likely to be made. If, for instance, the sponsor of the evaluation is the director of a local community mental health agency, then the evaluator may define the boundaries of the program around one of the distinct service packages administered by that director, such as outpatient counseling for eating disorders. If the evaluation sponsor is the state director of mental health, however, the relevant program boundaries may be defined around the outpatient counseling component of all the local mental health agencies in the state.

Because program theory deals mainly with means-ends relations, the most critical aspect of defining program boundaries is to ensure that they encompass all the important activities, events, and resources linked to one or more outcomes recognized as central to the endeavor. An evaluator accomplishes this by starting with the benefits the program intends to produce and working backward to identify and map all the organizational activities and resources presumed to contribute to attaining those objectives. From this perspective, the eating disorders program at either the local or state level would be defined as the set of activities organized by the respective mental health agency that has an identifiable role in attempting to alleviate eating disorders for the eligible population.

Although this approach is straightforward in concept, it can be problematic in practice. Not only can programs be complex, with crosscutting resources, activities, and goals, but the characteristics described above as linchpins for program definition can themselves be difficult to establish. Thus, in this matter, as with so many other aspects of evaluation, the evaluator must be prepared to negotiate a program definition agreeable to the evaluation sponsor and key stakeholders and be flexible about modifying the definition and resolving ambiguities in the program theory as the evaluation progresses.

Explicating the Program Theory

For a program in the early planning stage, program theory might be built by the planners from prior practice and research. At this stage, an evaluator may be able to help develop a plausible and well-articulated theory. For an existing program, however, the appropriate task is to describe the theory embodied in the program's structure and operation. To accomplish this, the evaluator must work with stakeholders to draw out the theory represented in their actions and assumptions. The general procedure for this involves successive approximation. Draft descriptions of the program theory are generated, usually by the evaluator, and discussed with knowledgeable stakeholder informants to get feedback. The draft is then refined on the basis of their input and shown again to appropriate stakeholders. The theory description developed in this fashion may involve impact theory, process theory, or any components or combination that are deemed relevant to the purposes of the evaluation. [Exhibit 3-J](#) presents an account of how a theory of action and a theory of change for a program designed to improve the performance of the lowest performing schools in North Carolina were elicited.

Exhibit 3-J Theory of Action and Theory of Change for Turning Around the Lowest Performing Schools in North Carolina

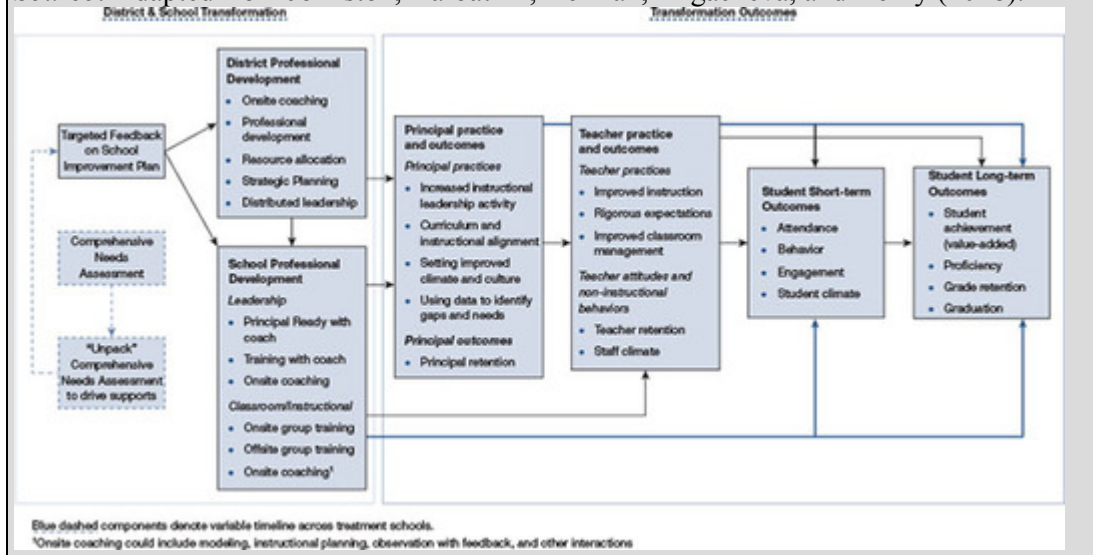
In 2015, the Department of Public Instruction in North Carolina prepared to initiate a new program to improve the performance of its lowest performing schools. The development of the program theory was based in part on documents from previous programs serving a similar purpose, in part on the conceptualization of the services needed by the leadership of the organizational units responsible for the services, and in part on legislative redefinition of what identified the lowest performing schools. The overall theory was divided into two distinct components: a theory of action, which described the activities undertaken by agency personnel to support the lowest performing schools (see the box labeled "District & School Transformation"), and a theory of change, which described the expected changes in the behaviors, attitudes, and skills of principals, teachers, and students.

Prior to delivery of the services, the theories of action and change were developed during the evaluation planning process through focus groups with agency leadership in which the evaluation team presented drafts of the theory to elicit reactions and proposed revisions. The depiction of the theory was revised several times by the evaluation team and resubmitted to the agency leadership until consensus on its representativeness was

achieved. After the services were initiated and before data collection for the evaluation began, the theory was refined once again to reflect the actual services being delivered.

The theory of action included many discrete services, such as a comprehensive needs assessment for each school and professional development for the principal and teachers. The theory of change then shows the expected direct effects of the services on principals and teachers, as well as the indirect effects on students' short-term and longer term outcomes.

Source: Adapted from Johnston, Harbatkin, Herman, Migacheva, and Henry (2018).



The primary sources of information for developing and differentiating descriptions of program theory are (a) review of program documents, (b) interviews with program stakeholders and other selected informants, (c) site visits and observation of program functions and circumstances, and (d) the social science literature. Three types of information the evaluator may be able to extract from those sources will be especially useful.

Program Goals and Objectives

Perhaps the most important matter to be determined from program sources relates to the goals and objectives of the program, which are necessarily an integral part of the program theory, especially its impact theory. The goals and objectives that must be represented in program theory, however, are not necessarily the same as those identified in a program's mission statements or in responses to questions asked of stakeholders. To be meaningful for an evaluation, program goals must identify a state of affairs that could

realistically be attained as a result of program actions; that is, there must be some reasonable connection between what the program does and what it intends to accomplish. To keep the discussion concrete and specific, the evaluator might use a line of questioning that does not ask about goals directly but asks instead about consequences. For instance, in a review of major program activities, the evaluator might ask about each, “Why do it? What are the expected results? How could you tell if those results actually occurred?”

The resulting set of goal statements must then be integrated into the description of program theory. Goals and objectives that describe the changes the program aims to bring about in social conditions relate to program impact theory. A program goal of reducing unemployment, for instance, identifies a distal outcome in the impact theory. Program goals and objectives related to program activities and service delivery, in turn, help reveal the program process theory. If the program aims to offer after-school programs for children who are not reading at grade level, a portion of the service utilization plan is revealed. Similarly, if an objective is to offer literacy classes four times a week, an important element of the organizational plan is identified.

Program Functions, Components, and Activities

To properly describe the program process theory, the evaluator must identify each distinct program component, its functions, and the particular activities and operations associated with those functions. Program functions include such operations as “assess client need,” “complete intake,” “assign case manager,” “recruit referral agencies,” “train field workers,” and the like. The evaluator can generally identify such functions by determining the activities and job descriptions of the various program personnel. When clustered into thematic groups, these functions represent the constituent elements of the program process theory.

The Logic or Sequence Linking Program Functions, Activities, and Components

A critical aspect of program theory is how the various expected outcomes and functions relate to each other. Sometimes these relationships involve only the temporal sequencing of key program activities and their effects; for instance, in a postrelease program for felons, prison officials must notify the program that a convict has been released before the program can initiate contact to arrange services. In other cases, the relationships between outcomes and functions have to do with activities or events that must be coordinated, as when child care and transportation must be arranged in conjunction with job training sessions, or with supportive functions, such as training the instructors who will conduct in-service classes for nurses. Other relationships entail logical or conceptual linkages, especially those represented in the program impact theory. For example, the connection between mothers' knowledge about how to care for their infants and the actual behavior of providing that care assumes a psychological process through which information influences behavior.

It is because the number and variety of such relationships are often appreciable that evaluators typically construct charts or graphical displays to describe them. These may be configured as lists, flowcharts, or hierarchies, or in any number of creative forms designed to identify the key elements and relationships in a program's theory. Such displays not only portray program theory but also provide a way to make it sufficiently concrete and specific to engage program personnel and stakeholders. Knowlton and Phillips (2013) provide numerous examples of creative displays of program theories that are tailored to the program and its organizational context, such as the Wayne Food Initiative program logic model, which uses a tree with roots and four branches that represent the four program strands (pp. 94-95, downloaded at <https://waynefoods.wordpress.com/home/program-logic-model/>).

Corroborating the Description of the Program Theory

The description of program theory that results from the procedures described will generally represent the program as it was intended more than as it actually is. Program managers and policymakers think of the idealized program as the real one with various shortfalls from that ideal as glitches that do not represent what the program is really about. Those further away from the day-to-day operations, on the other hand, may be unaware of such shortfalls and will naturally describe what they presume the program to be even if in actuality it does not quite live up to that image.

Some discrepancy between program theory and reality is therefore natural. Indeed, examination of the nature and magnitude of that discrepancy is the task of process or implementation evaluation, as discussed in the next chapter. However, if the theory is so overblown that it cannot realistically be held up as a depiction of what is supposed to happen, it needs to be revised. Suppose, for instance, that a job training program's service utilization plan calls for monthly contacts between each client and a case manager. If the program resources are insufficient to support case managers, and none are employed, this part of the theory is fanciful and should be restated to more realistically depict what the program might actually be able to accomplish.

In some cases, more nuanced ambiguities can arise in the corroboration of the program theory because of the use of terms that may not reflect a meaning shared by key stakeholders (Dahler-Larsen, 2017). For example, the theory of action presented in [Exhibit 3-J](#) includes on-site coaching. However, coaching had a more expansive definition for the coaches and organizational leadership than for some school personnel. In some instances, school personnel expressed disappointment that modeling instructional practices in an actual classroom or observing teachers and providing feedback were less frequent than expected. Dahler-Larsen raises even deeper ambiguity when describing “Janus variables”—variables that have a role in two different program theories. For example, coaching

provided by the state agency in the lowest performing schools may involve different approaches to instructional improvement and different expectations for instructional practices than the more general coaching provided by the school district. To manage expectations, it is important to develop clear, consensual definitions for the terms in a program theory and communicate them to the relevant stakeholders.

When the program theory depicts a realistic scenario, confirming it is a matter of demonstrating that pertinent program personnel and stakeholders endorse it as an adequate account of how the program is intended to work. If it is not possible to generate a theory description that all relevant stakeholders accept as applicable, this indicates that the program is poorly defined or that it embodies competing philosophies. In such cases, the most appropriate response for the evaluator may be to take on a consultant role and assist the program in clarifying its assumptions and intentions to yield a theory description that will be acceptable to all key stakeholders.

For the evaluator, the end result of the theory description exercise is a detailed and complete statement of the program as intended that can then be analyzed and assessed as a distinct form of evaluation. Note that the agreement of stakeholders serves only to confirm that the theory description does in fact represent their understanding of how the program is supposed to work. It does not necessarily mean that the theory is a good one. To determine the soundness of a program theory, the evaluator must not only describe the theory but evaluate it. The procedures evaluators use for that purpose are described in the next section.

Assessing Program Theory

Assessment of some aspect of a program's theory is relatively common in evaluation, often in conjunction with an evaluation of program process or impact. Nonetheless, outside of the evaluability assessment literature, remarkably little has been written about how this should be done. Our interpretation of this relative neglect is not that theory assessment is unimportant or unusual, but that it is typically done in an informal manner that relies on commonsense judgments that may not seem to require much explanation or justification. Indeed, when program services are directly related to straightforward objectives, the validity of the program theory may be accepted on the basis of limited evidence or commonsense judgment. An illustration is a meals-on-wheels service that brings hot meals to homebound elderly persons to improve their nutritional intake. In this case, the theory linking the action of the program (providing hot meals) to its intended benefits (improved nutrition) needs little critical evaluation.

Many programs, however, are not based on expectations as simple as the notion that delivering food to elderly persons improves their nutrition. For example, a family preservation program that assigns case managers to coordinate community services for parents deemed at risk of having their children placed in foster care involves many assumptions about exactly what it is supposed to accomplish and how. In such cases, the program theory might easily be faulty, and correspondingly, a rather probing evaluation of it may be warranted.

It is seldom possible or useful to individually appraise each distinct assumption and expectation represented in a program theory. But there are certain critical tests that can be conducted to provide assurance that it is sound. This section summarizes the various approaches and procedures the evaluator might use for conducting that assessment.

Assessment in Relation to Social Needs

The most important framework for assessing program theory builds on the results of needs assessment, as discussed in [Chapter 2](#). Or, more generally, it is based on a thorough understanding of the social problem the program is intended to address and the service needs of the target population. A program theory that does not relate in an appropriate manner to the actual nature and circumstances of the social conditions at issue will result in an ineffective program no matter how well the program is implemented and administered. It is fundamental, therefore, to assess program theory in relationship to the needs of the target population the program is intended to serve.

There is no push-button procedure an evaluator can use to assess whether program theory describes a suitable conceptualization of how social needs should be met. Inevitably, this assessment requires judgment calls. When the assessment is especially critical, its validity is strengthened if those judgments are made collaboratively with relevant experts and stakeholders to broaden the range of perspectives and expertise on which they are based. Such collaborators, for instance, might include social scientists knowledgeable about research and theory related to the intervention, administrators with long experience managing such programs, representatives of advocacy groups associated with the target population, and policymakers or policy advisers familiar with the program and problem area.

Whatever the nature of the group that contributes to the assessment, the crucial aspect of the process is specificity. When program theory and social needs are described in general terms, there often appears to be more correspondence than is evident when the details are examined. To illustrate, consider a curfew program prohibiting juveniles under age 18 from being outside their homes after midnight that is initiated in a metropolitan area to address the problem of skyrocketing juvenile crime. The program theory, in general terms, is that the curfew will keep youths home at night, and if they are at home, they are unlikely to commit crimes. Because the general social problem the program addresses is juvenile crime, the program theory does seem responsive to the social need.

A more detailed problem diagnosis and service needs assessment, however, might show that the bulk of juvenile crimes are residential burglaries committed in the late afternoon when school lets out. Moreover, it might reveal that the offenders represent a relatively small proportion of the juvenile population who have a disproportionately large impact because of their high rates of offending. Furthermore, it might be found that these juveniles are predominantly youths who have no supervision during after-school hours. When the program theory is then examined in some detail, it is apparent that it assumes that significant juvenile crime occurs late at night and that potential offenders will both know about and obey the curfew. Furthermore, it depends on enforcement by parents or the police if compliance does not occur voluntarily.

Although even more specificity than this would be desirable, this much detail illustrates how a program theory can be compared with problem diagnosis and the need to discover shortcomings in the theory. In this example, examining the particulars of the program theory and the social problem it is intended to address reveals a large disconnect. The program blankets the whole city rather than targeting the small group of problem juveniles and focuses on activity late at night rather than during the early afternoon, when most of the crimes actually occur. In addition, it makes the questionable assumptions that youths already engaged in more serious lawbreaking will comply with a curfew, that parents who leave their delinquent children unsupervised during the early part of the day will be able to supervise their later behavior, and that the overburdened police force will invest sufficient effort in arresting juveniles who violate the curfew to enforce compliance. Careful review of these particulars alone would raise serious doubts about the validity of this program theory.

One useful approach to comparing program theory with what is known (or assumed) about the relevant social needs is to separately assess impact theory and program process theory. Each of these relates to the social problem in a different way and, as each is elaborated, specific questions can be asked about how compatible the assumptions of the theory are with the nature of the social circumstances to which it applies. We will briefly describe the main points of comparison for each of these theory components.

Program impact theory involves the sequence of causal links between program services and outcomes that improve the targeted social conditions. The key point of comparison between program impact theory and social needs, therefore, relates to whether the effects the program is expected to have on the social conditions correspond to what is required to improve those conditions, as revealed by the needs assessment. Consider, for instance, a school-based educational program aimed at getting elementary school children to learn and practice good eating habits. The problem this program attempts to ameliorate is poor nutritional choices among school-age children, especially those in economically disadvantaged areas. The program impact theory would show a sequence of links between the planned instructional exercises and the children's awareness of the nutritional value of foods, culminating in healthier selections and therefore improved nutrition.

Now, suppose a thorough needs assessment shows that the children's eating habits are indeed poor but that their nutritional knowledge is not especially deficient. The needs assessment further shows that the foods served at home and even those offered in the school cafeterias provide limited opportunity for healthy selections. Against this background, it is evident that the program impact theory is flawed. Even if the program successfully imparts additional information about healthy eating, the children will not be able to act on it because they have little control over the selection of foods available to them. Thus, the proximal outcomes the program impact theory describes may be achieved, but they are not what is needed to ameliorate the problem at issue.

Program process theory, on the other hand, represents assumptions about the capability of the program to provide services that are accessible to the target population and compatible with their needs. These assumptions, in turn, can be compared with information about the target population's opportunities to obtain service and the barriers that inhibit them from using the service. The process theory for an adult literacy program that offers evening classes at the local high school, for instance, may incorporate instructional and advertising functions and an appropriate selection of courses for the target population. The details of this scheme can be compared with needs assessment data that show what logistical and

psychological support the target population requires to make effective use of the program. Child care and transportation may be critical for some potential participants. Also, illiterate adults may be reluctant to enroll in courses without more personal encouragement than they would receive from advertising. Cultural and personal affinity with the instructors may be important factors in attracting and maintaining participation from the target population as well. The intended program process can thus be assessed in terms of how responsive it is to these dimensions of the needs of the target population.

Assessment of Logic and Plausibility

A thorough job of articulating program theory should reveal the critical assumptions and expectations inherent in the program's design. One essential form of assessment is simply a critical review of the logic and plausibility of these aspects of the program theory. Commentators familiar with assessing program theory suggest that a panel of reviewers be organized for that purpose (Chen, 1990; Rutman, 1980; Smith, 1989; Wholey, 2015). Such an expert review panel should include representatives of the program staff and other major stakeholders as well as the evaluator. By definition, however, stakeholders have some direct stake in the program. To balance the assessment and expand the available expertise, it will be advisable to bring in informed persons with no direct relationship to the program. Such outside experts might include experienced administrators of similar programs, social researchers with relevant specialties, representatives of advocacy groups or client organizations, and the like.

Exhibit 3-K GREAT Program Theory Is Consistent With Criminological Research

In 1991 the Phoenix, Arizona, Police Department initiated a program with local educators to provide youths in the elementary grades with the tools necessary to resist becoming gang members. Known as GREAT (Gang Resistance Education and Training), the program has attracted federal funding and is now distributed nationally. The program is taught to seventh graders in schools over 9 consecutive weeks by uniformed police officers. It is structured around detailed lesson plans that emphasize teaching youths how to set goals for themselves, how to resist peer pressure, how to resolve conflicts, and how gangs can affect the quality of their lives.

The program has no officially stated theoretical grounding other than Glasser's (1975) reality therapy, but GREAT training officers and others associated with the program make reference to sociological and psychological concepts as they train GREAT instructors. As part of an analysis of the program's impact theory, a team of criminal justice researchers identified two well-researched criminological theories relevant to gang participation: Gottfredson and Hirschi's self-control theory (SCT) and Akers's social learning theory (SLT). They then reviewed the GREAT lesson plans to assess their consistency with the most pertinent aspects of these theories. To illustrate their findings, a summary of Lesson 4 is provided below, with the researchers' analysis in *italics* after the lesson description:

Lesson 4. Conflict Resolution: Students learn how to create an atmosphere of understanding that would enable all parties to better address problems and work on solutions together. *This lesson includes concepts related to SCT's anger and aggressive coping strategies. SLT ideas are also present: Instructors present peaceful, nonconfrontational means of resolving conflicts. Part of this lesson deals*

with giving the student a means of dealing with peer pressure to join gangs and a means of avoiding negative peers with a focus on the positive results (reinforcements) of resolving disagreements by means other than violence. Many of these ideas directly reflect constructs used in previous research on social learning and gangs.

Similar comparisons showed good consistency between the concepts of the criminological theories and the lesson plans for all but one of the eight lessons. The reviewers concluded that the GREAT curriculum contained implicit and explicit linkages both to SCT and SLT.

Source: Adapted from Winfree, Esbensen, and Osgood (1996).

A review of the logic and plausibility of program theory will necessarily be a relatively unstructured and open-ended process. Nonetheless, there are some general issues such reviews should address. These are described below in the form of questions reviewers can ask. Additional useful detail can be found in Rutman (1980), Smith (1989), and Wholey (2015).

- Are the program goals and objectives well defined? The outcomes for which the program is accountable should be stated in sufficiently clear and concrete terms to permit a determination of whether they have been attained. Goals such as “introducing students to computer technology” are not well defined in this sense, whereas “increasing student knowledge of the ways computers can be used” is well defined and measurable.
- Are the program goals and objectives feasible? That is, is it realistic to assume that they can actually be attained as a result of the services the program delivers? A program theory should specify expected outcomes that are of a nature and scope that might reasonably follow from a successful program and that do not represent unrealistically high expectations. Moreover, the stated goals and objectives should involve conditions the program might actually be able to affect in some meaningful fashion, not those largely beyond its influence. “Eliminating poverty” is grandiose for any program, whereas “decreasing the unemployment rate” is not. But even the latter goal might be unrealistic for a job training program that can enroll only 50 students at a time.
- Is the change process assumed in the program theory plausible? The presumption that a program will create benefits for the intended target

population depends on the occurrence of some cause-and-effect chain that begins with the targets' interaction with the program and ends with the improved circumstances in the target population that the program expects to bring about. Every step of this causal chain should be plausible. Because the validity of this impact theory is the key to the program's ability to produce the intended effects, it is best if the theory is supported by evidence that the assumed links and relationships actually occur. For example, suppose a program is based on the presumption that exposure to literature about the health hazards of drug abuse will motivate long-term heroin addicts to renounce drug use. In this case, the program theory does not present a plausible change process, nor is it supported by any research evidence.

- Are the procedures for identifying members of the target population, delivering service to them, and sustaining that service through completion well defined and sufficient? The program theory should specify procedures and functions that are both well defined and adequate for the purpose, viewed both from the perspective of the program's ability to perform them and the target population's likelihood of being engaged by them. Consider, for example, a program to test for high blood pressure among poor and elderly populations to identify those needing medical care. It is relevant to ask whether this service is provided in locations accessible to members of these groups and whether there is an effective means of locating those with uncertain addresses. Absent these characteristics, it is unlikely that many persons from the target groups will receive the intended service.
- Are the constituent components, activities, and functions of the program well defined and sufficient? A program's structure and process should be specific enough to permit orderly operations, effective management control, and monitoring by means of attainable, meaningful performance measures. Most critical, the program components and activities should be sufficient and appropriate to attain the intended goals and objectives. A function such as "client advocacy" has little practical significance if no personnel are assigned to it or there is no common understanding of what it means operationally. A relatively recent approach for addressing this question is drill-down logic model review that specifies and sequences the

activities needed to produce each program output and achieve its objectives (Peyton & Scicchitano, 2017). The process begins with a review or development of an initial logic model, gathering information from documents and interviews about how the program actually operates, revising the logic model, and then developing more detailed submodels that include the sequence of well-defined steps in the process for each output in the model.

- Are the resources allocated to the program and its various activities adequate? Program resources include not only funding but also personnel, material, equipment, facilities, relationships, reputation, and other such assets. There should be a reasonable correspondence between the program as described in the program theory and the resources available for operating it. A program theory that calls for activities and outcomes that are unrealistic relative to available resources cannot be said to be a good theory. For example, a management training program too short staffed to initiate more than a few brief workshops cannot expect to have a significant impact on management skills in the organization.

Assessment Through Comparison With Research and Practice

Although every program is distinctive in some ways, few are based entirely on unique assumptions about how to engender change, deliver service, and perform major program functions. Some information applicable to assessing the various components of program theory is likely to exist in the social science and human services research literature. One useful approach to assessing program theory, therefore, is to find out whether it is congruent with research evidence and practical experience elsewhere ([Exhibit 3-K](#) summarizes one example of this approach).

There are several ways in which evaluators might compare a program theory with findings from research and practice. The most straightforward is to examine evaluations of programs based on similar concepts. The results will give some indication of the likelihood that a program will be successful and perhaps identify critical problem areas. Evaluations of very similar programs, of course, will be the most informative in this regard. However, evaluation results for programs that are similar only in terms of general theory, even if different in other regards, might also be instructive.

Consider a mass media campaign in a metropolitan area to encourage women to have mammographic screening for early detection of breast cancer. The impact theory for this program presumes that exposure to TV, radio, and newspaper messages will stimulate a reaction that will result in increased rates of screening. The credibility of the impact theory assumed to link exposure and increases in testing is enhanced by evidence that similar media campaigns in other cities have resulted in increased mammographic testing. Moreover, the program's process theory also gains some support if the evaluations for other campaigns show that the program functions and scheme for delivering messages to the target population were similar to that intended for the program at issue. Suppose, however, that no evaluation results are available about media campaigns promoting mammographic screening in other cities. It might still be informative to examine information about analogous media campaigns. For instance,

reports may be available about media campaigns to promote immunizations, dental checkups, or other such actions that are health related and require a visit to a provider. So long as these campaigns involve similar principles, their success might well be relevant to assessing the program theory on which the mammography campaign is based.

In some instances, basic research on the social and psychological processes central to the program may be available as a framework for assessing the program theory, particularly impact theory. Unfortunately for the evaluation field, relatively little basic research has been done on the social dynamics that are common and important to intervention programs. Where such research exists, however, it can be very useful. For instance, a mass media campaign to encourage mammographic screening involves messages intended to change attitudes and behavior. The large body of basic research in social psychology on attitude change and its relationship to behavior provides some basis for assessing the impact theory for such a media campaign. One established finding is that messages designed to raise fears are generally less effective than those providing positive reasons for a behavior. Thus, an impact theory based on the presumption that increasing awareness of the dangers of breast cancer will prompt increased mammographic screening may not be a good one.

There is also a large applied research literature on media campaigns and related approaches in the field of advertising and marketing. Although this literature largely has to do with selling products and services, it too may provide some basis for assessing the program theory for the breast cancer media campaign. Market segmentation studies, for instance, may show what media and what times of the day are best for reaching women with various demographic profiles. The evaluator can then use this information to examine whether the program's service utilization plan is optimal for communicating with women whose age and circumstances put them at risk for breast cancer.

Use of the research literature to help with assessment of program theory is not limited to situations of good overall correspondence between the programs or processes the evaluator is investigating and those represented in the research. An alternate approach is to break the theory down into its

component parts and linkages and search for research evidence relevant to each component. Much of program theory can be stated as “if-then” propositions: If case managers are assigned, then more services will be provided; if school performance improves, then delinquent behavior will decrease; if teacher-to-student ratios are higher, then students will receive more individual attention. Research may be available that indicates the plausibility of individual propositions of this sort. The results, in turn, can provide a basis for a broader assessment of the theory with the added advantage of identifying any especially weak links. This approach was pioneered by the Program Evaluation and Methodology Division of the U.S. General Accounting Office as a way to provide rapid review of program proposals arising in the Congress (Cordray, 1993; U.S. General Accounting Office, 1990).

Assessment via Preliminary Observation

Program theory, of course, is inherently conceptual and cannot be observed directly. Nonetheless, it involves many assumptions about how things are supposed to work that an evaluator can assess by observing the program in operation, talking to staff and service recipients, and making other such inquiries focused specifically on the program theory. Indeed, a thorough assessment of program theory of programs that are in operation should incorporate some firsthand observation and not rely entirely on logical analysis and armchair reviews. Direct observation provides a reality check on the concordance between program theory and the program it is supposed to describe. Consider a program for which it is assumed that distributing brochures about good nutrition to senior citizens centers will influence the eating behavior of persons over age 65. Observations revealing that the brochures are rarely read by anyone attending the centers would certainly raise a question about the assumption that the target population will be exposed to the information in the brochures, a precondition for any attitude or behavior change.

To assess a program's impact theory, the evaluator might conduct observations and interviews focusing on the participant-program interactions that are expected to produce the intended outcomes. This inquiry would look into whether those outcomes are appropriate for the program circumstances and whether they are realistically attainable. For example, consider the presumption that a welfare-to-work program can enable a large proportion of welfare clients to find and maintain employment. To gauge how realistic the intended program outcomes are, the evaluator might examine the local job market, the work readiness of the welfare population (number physically and mentally fit, skill levels, work histories, motivation), and the economic benefits of working relative to staying on welfare. At the service end of the change process, the evaluator might observe job training activities and conduct interviews with participants to assess the likelihood that the intended changes would occur.

To test the service utilization component of a program's process theory, the evaluator could examine the circumstances of the target population to better understand how and why they might become engaged with the program.

This information would permit an assessment of the quality of the program's service delivery plan for locating, recruiting, and serving the intended clientele. To assess the service utilization plan of a midnight basketball program to reduce delinquency among high-risk youths, for instance, the evaluator might observe the program activities and interview participants, program staff, and neighborhood youths about who participates and how regularly. The program's service utilization assumptions would be supported by indications that the most delinquency-prone youths participate regularly in the program.

Finally, the evaluator might assess the plausibility of the organizational component of the program's process theory through observations and interviews relating to program activities and the supporting resources. Critical here is evidence that the program can actually perform the intended functions. Consider, for instance, a program plan that calls for sixth grade science teachers throughout a school district to take their students on two science-related field trips per year. The evaluator could probe the presumption that this would actually be done by interviewing a number of teachers and principals to find out the feasibility of scheduling, the availability of buses and funding, and the like.

Note that any assessment of program theory that involves collection of new data could easily turn into a full-scale investigation of whether what was presumed in the theory actually happened. Here, however, our focus is on the task of assessing the soundness of the program theory description as a plan, that is, as a statement of the program as intended rather than as a statement of what is actually happening (that assessment comes later). In recognizing the role of observation and interview in the process, we are not suggesting that theory assessment necessarily requires a full evaluation of the program. Instead, we are suggesting that some appropriately configured contact with the program activities, target population, and related situations and informants can provide the evaluator with valuable information about how plausible and realistic the program theory is.

Possible Outcomes of Program Theory Assessment

A program whose design is weak or faulty has little prospect for success even if it adequately implements that design. Thus, if the program theory is not sound, there may be little reason to assess other evaluation issues, such as the program's implementation, impact, or efficiency. Within the framework of evaluability assessment, finding that the program theory is poorly defined or seriously flawed indicates that the program simply is not yet evaluable.

When assessment of program theory reveals deficiencies, one appropriate response is for the responsible parties to redesign the program. Such program reconceptualization may include (a) clarifying goals and objectives; (b) restructuring program components for which the intended activities are not happening, needed, or reasonable; and (c) working with stakeholders to obtain consensus about the logic that connects program activities with the desired outcomes. The evaluator may guide or facilitate this process.

If an evaluation of program process or impact goes forward without articulation of a credible program theory, then a certain amount of ambiguity will be inherent in the results. This ambiguity is potentially twofold. First, if program process theory is not well defined, there is ambiguity about what the program is expected to be doing operationally. This complicates the identification of criteria for judging how well the program is implemented. Such criteria must then be established individually for the various key program functions through some piecemeal process. For instance, administrative criteria may be stipulated regarding the number of clients to serve, the amount of service to provide, and the like, but they will not be integrated into an overall plan for the program.

Second, if there is no adequate specification of the program impact theory, an impact evaluation may be able to determine whether certain outcomes were produced (see [Chapters 6](#) to [8](#)), but it will be difficult to explain why

or—often more important—why not. Poorly specified impact theory limits the ability to identify or measure the intervening variables on which the outcomes may depend and, correspondingly, the ability to explain what went right or wrong in producing the expected outcomes. If program process theory is also poorly specified, it will not even be possible to adequately describe the nature of the program that produced, or failed to produce, the outcomes of interest. Evaluation under these circumstances is often referred to as **black-box evaluation** to indicate that assessment of outcomes is made without much insight into what is causing those outcomes.

Only a well-defined and well-justified program theory permits ready identification of critical program functions and what is supposed to happen as a result. This structure provides meaningful benchmarks against which both managers and evaluators can compare actual program performance. The framework of program theory, therefore, gives the program a blueprint for effective management and gives the evaluator guidance for designing the process, impact, and efficiency evaluations described in subsequent chapters.

Summary

- Program theory is an aspect of a program that can be evaluated in its own right. Such assessment is important because a program based on a weak or faulty conceptualization has little prospect of achieving the intended results.
- The most fully developed approaches to evaluating program theory have been described in the context of evaluability assessment, an appraisal of whether a program's performance can be evaluated and, if so, whether it should be. Evaluability assessment involves describing program goals and objectives, assessing whether the program is well enough conceptualized to be evaluable, and identifying stakeholder interest in using evaluation findings.
- Evaluability assessment may result in efforts by program managers to better conceptualize their program. It may indicate that the program is too poorly defined for evaluation or that there is little likelihood that the findings will be used. Alternatively, it could find that the program theory is well defined and plausible, that evaluation findings will likely be used, and that a meaningful evaluation could be done.
- To assess program theory, it is first necessary for the evaluator to describe the theory in a clear, explicit form acceptable to stakeholders. The aim of this effort is to describe the "program as intended" and its rationale, not the program as it actually is. Three key components that should be included in this description are the program impact theory, the service utilization plan, and the program's organizational plan.

- The assumptions and expectations that make up a program theory may be well formulated and explicitly stated (thus constituting an articulated program theory), or they may be inherent in the program but not overtly stated (thus constituting an implicit program theory). When a program theory is implicit, the evaluator must extract and articulate the theory by collating and integrating information from program documents, interviews with program personnel and other stakeholders, and observations of program activities.
- When articulating an implicit program theory, it is especially important to formulate clear, concrete statements of the program's goals and objectives as well as an account of how the desired outcomes are expected to result from program action. The evaluator should seek corroboration from stakeholders that the resulting description meaningfully and accurately describes the "program as intended."
- There are several approaches to assessing program theory. The most important assessment the evaluator can make is based on a comparison of the intervention specified in the program theory with the social needs the program is expected to address. Examining critical details of the program conceptualization in relation to the social problem indicates whether the program represents a reasonable plan for ameliorating that problem. This analysis is facilitated when a needs assessment has been conducted to systematically diagnose the problematic social conditions ([Chapter 2](#)).
- A complementary approach to assessing program theory uses stakeholders and other informants to appraise the clarity, plausibility, feasibility, and appropriateness of the program theory as formulated.
- Program theory can also be assessed in relation to the support for its critical assumptions found in research or documented practice elsewhere. Sometimes findings are available for similar programs, or programs based on similar theory, so that the evaluator can make an overall comparison between a program's theory and relevant evidence. If the research and practice literature does not support overall comparisons, however, evidence bearing on specific key relationships assumed in the program theory may still be obtainable.
- Evaluators can often usefully supplement other approaches to assessment with direct observations to further probe critical assumptions in the program theory.
- Assessment of program theory may indicate that the program is not evaluable because of basic flaws in its theory. Such findings are an important evaluation product in their own right and can be informative for program stakeholders. In such cases, one appropriate response is to redesign the program, a process that the evaluator may guide or facilitate.
- If evaluation of program process or impact proceeds without articulation of a credible program theory, the results will be ambiguous. In contrast, a sound program theory provides a basis for evaluation of how well that theory is implemented, what effects are produced on the target outcomes, and how efficiently they are produced—topics to be discussed in subsequent chapters.

Key Concepts

Articulated program theory 71

Black-box evaluation 87

Evaluability assessment 61

Impact theory 65

Implicit program theory 74

Organizational plan 67

Process theory 67

Service utilization plan 67

Critical Thinking/Discussion Questions

1. Describe the three primary activities in an evaluability assessment. What is the expected outcome of an evaluability assessment, and what is its overarching purpose?
2. Explain the three components of program theory—the program impact theory, the service utilization plan, and the program’s organizational plan—and describe how they are interrelated.
3. There are several ways in which evaluators might compare a program theory with findings from research and practice. Explain three ways in which this can be done and provide examples.

Application Exercises

1. Choose a social program you are familiar with. Review its Web site and any organizational materials you can access and prepare a logic model for the program. Be sure to include inputs, outputs, and outcomes. Explain how you think the proximal outcomes are related to the distal outcomes.
2. Locate an evaluation report that discusses program theory. First describe the program that was evaluated. Then discuss how the program theory was developed. Was the program theory implicit or explicit? How complete do you think the program theory is in relation to the description of the elements of program theory presented in this chapter?

Chapter 4 Assessing Program Process and Implementation

What Is Process Evaluation and Monitoring?

Setting Criteria for Judging Program Process

Common Forms of Process Evaluations

Process Evaluation

Process Monitoring and Administrative Data Systems

Perspectives on Program Process Monitoring

Process Assessment From the Evaluator's Perspective

Process Assessment From an Accountability Perspective

Process Assessment From a Management Perspective

Assessing Service Utilization

Coverage and Bias

Measuring Coverage

Program Records

Surveys

Assessing Bias: Program Users, Eligibles, and Dropouts

Assessing Organizational Functions

The Delivery System

Specification of Services

Accessibility

Program Support Functions

Summary

Key Concepts

To be effective in bringing about the desired improvements in social conditions, a program needs more than a good design. The program staff also must implement its design; that is, it must actually carry out its intended functions in the intended way.

Although implementing a program concept may seem straightforward, in practice it is often difficult. Social programs typically must contend with many adverse influences that can compromise even well-intentioned attempts to conduct program business appropriately. The result can easily be substantial discrepancies between the program as intended and the program as actually implemented.

The implementation of a program is reflected in concrete form in the program processes that it puts in place. An important evaluation function, therefore, is to assess the adequacy of program process: the program activities that actually take place and the services that are actually delivered in routine program operation. A related function is to examine the fidelity of implementation: the extent to which the services are consistent with the design of the program. When process evaluation occurs on an ongoing, periodic basis, it is referred to as process monitoring. This chapter introduces the procedures evaluators use to investigate these issues.

In this chapter, we return to a theme in previous chapters: A solid design for a social program that is built on an accurate understanding of the needs of the program's target population is not enough to ensure that the desired outcomes will be achieved. The program must be implemented in a manner consistent with the design and delivered with sufficient quality, frequency, and intensity to the targeted beneficiaries to realize the intended benefits if its theory of change is valid. Many steps are required to take a program from concept to full operation, and much effort is needed to keep it true to its original design and purposes. Thus, whether any program is fully carried out as envisioned by its sponsors and managers is always an appropriate topic for systematic evaluation.

Ascertaining how well a program is operating, therefore, is an important and useful form of evaluation, known as *process evaluation*. Process evaluation does not represent a single distinct evaluation procedure but, rather, a family of approaches, concepts, and methods. The defining theme of process evaluation is a focus on the enacted program itself: its operations, activities, functions, performance, staffing, resources, and so forth. When process evaluation involves an ongoing effort to measure and record information about the program's operation, we will refer to it as *process monitoring*. When the process evaluation focuses on the consistency of program operations with the design of the program, it is referred to as *implementation fidelity*. In some fields, including public health and international development, it is common to fold the monitoring of program processes into a broader set of evaluative activities known as *monitoring and evaluation*, or M&E. **Monitoring and evaluation** is the practice of ongoing collection and reporting of data on program activities, products, and outcomes along with resource utilization and staffing for managing the program combined with outcome or impact evaluation at appropriate points in the life cycle of the program.

What Is Program Process Evaluation and Monitoring?

Evaluators distinguish between process evaluation and impact evaluation. **Process evaluation** examines what a program is, the activities undertaken, who receives services or other benefits, and the consistency with which it is implemented in terms of its design and across sites. Often it is undertaken for formative or program improvement purposes: It can directly point to deficiencies in the ongoing operations of a program that may be remedied by its administrators. Also, it can be a crucial element in interpreting effect estimates from impact evaluations. It does not, however, attempt to assess the effects of the program on its recipients. Such assessment is the province of impact evaluation. **Process monitoring** is the systematic, periodic documentation of key aspects of program performance that assesses whether the program is operating as intended or according to some appropriate standard. By parallel construction, **outcome monitoring** is the periodic measurement of the outcomes of interest to the program on the program participants.

Program process evaluation generally involves assessments of program performance in the domains of service utilization and program organization. Assessing service utilization consists of examining the extent to which the intended target population receives the intended services. Assessing program organization requires comparing the plan for what the program should be doing with what is actually done, especially with regard to providing services. Usually, process evaluation is directed at one or both of two key questions: (a) whether a program is reaching the appropriate target population and (b) whether its service delivery and support functions are consistent with the program design specifications or other appropriate standards. More specifically, process evaluation is designed to answer such evaluation questions as these:

- How many persons are receiving services?
- Are those receiving services members of the intended target population?

- Are they receiving the proper amount, type, and quality of services?
- Are there members of the target population who are not receiving services or subgroups within that population who are underrepresented among those receiving services?
- Are members of the target population aware of the program?
- Are necessary program functions being performed adequately?
- Is staffing sufficient in numbers and qualifications for the functions that must be performed?
- Is the program well organized? Do staff work well with one another?
- Does the program coordinate effectively with the other programs and agencies with which it must interact?
- Are resources, facilities, and funding adequate to support necessary program functions?
- Are resources used effectively and efficiently?
- Is the program implemented as designed?
- Does the program comply with requirements imposed by its governing board, funding agencies, or higher level administration?
- Does the program comply with applicable professional and legal standards?
- Do program operations or performance vary significantly between sites or locales?
- Are participants satisfied with their interactions with program personnel and procedures?
- Are participants satisfied with the services they receive?
- Do participants engage in appropriate follow-up behavior after service?

Setting Criteria for Judging Program Process

It is important to recognize the evaluative aspects of process evaluation questions such as those listed above. Virtually all those questions involve words such as *appropriate*, *adequate*, *sufficient*, *satisfactory*, *reasonable*, *intended*, and other phrasing indicating that an evaluative judgment is required. To answer these questions, therefore, the evaluator or other responsible parties must not only describe the program's performance but also assess whether it is satisfactory. This, in turn, requires that there be some bases for making judgments, that is, some defensible criteria or standards to apply. Where such criteria are not already articulated and endorsed, the evaluator may find that establishing workable criteria is as difficult as measuring program performance on the pertinent dimensions.

There are several approaches to setting criteria for program performance. Moreover, different approaches will apply to different dimensions of program performance because the considerations that go into defining, say, what constitutes an appropriate number of clients served are different from those pertinent to deciding whether the service personnel are providing an adequate quality of service. This said, the approach to the criterion issue that has the broadest scope and most general utility in process evaluation is the application of program theory as described in [Chapter 3](#).

Recall that program theory, as we presented it, is divided into program process theory and program impact theory. Program process theory is formulated to describe the program as intended in a form that virtually constitutes a plan or blueprint for what the program is expected to do and how. As such, it is particularly relevant to program process evaluation. Recall also that program theory builds on needs assessment (whether systematic or informal) and thus connects the program design with the social conditions the program is intended to ameliorate. And, of course, the process through which theory is derived and adopted usually involves input from major stakeholders and, ultimately, their endorsement. Program theory thus has a certain authority in delineating what a program "should" be doing and, correspondingly, what constitutes adequate performance.

Process evaluation, therefore, can be built on the foundation of program process theory. Process theory identifies the aspects of program performance most important to describe and also provides some indication of what level of performance is intended, thereby providing the basis for assessing whether actual performance measures up. [Exhibit 3-F](#) in the previous chapter, for instance, illustrates the service utilization component of the program process theory for an aftercare program for released psychiatric patients. That flowchart depicts, step by step, the interactions and experiences patients released from the hospital are supposed to have as a result of program service. A thorough process evaluation would systematically document what actually happened at each step. In particular, it would, for example, report how many patients were released from the hospital each month, what proportion were visited by a social worker, how many were referred to services and which services, and how many actually received those services.

If the program processes that are supposed to happen do not happen, then we would judge the program's performance to be poor. In actuality, of course, the situation is rarely so simple. Most often, critical events will not occur in an all-or-none fashion, but will be attained to some higher or lower degree. Thus, some, but not all, of the released patients will receive visits from social workers, some will be referred to services, and so forth. Moreover, there may be important quality dimensions. For instance, it would not represent good program performance if a released patient were referred to several community services, but these services were inappropriate to the patient's needs. To determine how much must be done, or how well, additional criteria are needed that parallel the information the process data provide. If the process data show that 63% of the released patients are visited by a social worker within 2 weeks of release, we cannot evaluate that performance without some standard that tells us what percentage is "good." Is 63% a poor performance, given that we might expect 100% to be desirable, or is it a very impressive performance with a clientele that is difficult to locate and serve?

The most common and widely applicable criteria for such situations are simply **administrative standards** or objectives, that is, stipulated target achievement levels set by program administrators or other responsible

parties. For example, the director and staff of a job training program may commit to attaining 80% completion rates for the training or to having 60% of the participants employed in stable positions 6 months after receiving training. For the psychiatric aftercare program, the administrative target might be to have 75% of the patients visited within 2 weeks of release from the hospital. By this standard, 63% is a subpar performance that, nonetheless, is not too far below the mark.

Administrative standards and objectives for program process performance may be set on the basis of past experience, the performance of comparable programs (often referred to as *benchmarking*), or simply the professional judgment of program managers or advisers. If they are reasonably justified, administrative standards can provide meaningful criteria for assessing observed program performance. In a related vein, some aspects of program performance may fall under applicable legal, ethical, or professional standards. The standards of care adopted in medical practice for treating common ailments, for instance, provide a set of criteria against which to assess program performance in health care settings. Similarly, state children's protective services typically have legal requirements to meet concerning handling cases of possible child abuse or neglect.

In practice, the assessment of particular dimensions of program process performance is often not based on specific, predetermined criteria but represents an after-the-fact judgment call. This is the "I'll know it when I see it" school of thought on what constitutes good program performance. An evaluator who collects process data on, say, the proportion of high-risk adolescents who recall seeing program-sponsored antidrug media messages may find program staff and other key stakeholders resistant to stating what an acceptable proportion would be. If the results come in at 50%, however, a consensus may arise that this is rather good considering the nature of the population, even though some stakeholders might have reported much higher expectations prior to seeing the data. Other findings, such as 40% or 60%, might also be considered rather good. Only extreme findings, say 10%, might strike all stakeholders as distressingly low. In short, without specific prior criteria, a wide range of performance might be regarded as acceptable. Of course, assessment procedures that are too flexible and that lead to a "pass" for all tend to be useless.

Some program designs call for tailoring the services to particular individuals or other units, such as schools or clinics. Tailoring services to the needs of the client or service unit complicates the determination of appropriate standards for judging the adequacy or sufficiency of the services. For example, in the process and implementation evaluation of the program to improve the lowest performing schools in North Carolina, depicted in [Exhibit 3-J](#) in the previous chapter, a question arose about the adequate amount of coaching for principals. An evaluation documented that between January 2016 and June 2017, the coaches completed a total of 1,896 visits to schools, ranging from 6 to 63 visits across schools. The tailored nature of the coaching made it difficult to judge if 63 visits was too many or 6 was too few over the 18 months. Rather than basing the assessment on the number of visits, the evaluation team surveyed principals, asking whether they met with their coaches regularly and if they viewed the amount of coaching as sufficient to meet the needs of the schools and their needs as school leaders. The responses indicated that 73% of the principals believed the intensity of the coaching met their needs, a figure judged by an expert advisory panel to represent acceptable performance.

Very similar considerations apply to the organizational component of program process theory. A depiction of the organizational plan for the psychiatric aftercare program was presented in [Exhibit 3-G](#) in [Chapter 3](#). Looking back at it will reveal that it too identifies dimensions of program performance that can be described and assessed against appropriate standards. Under that plan, for instance, case managers are expected to interview clients and families, assess service needs, and make referrals to services. A program process evaluation would document and assess what was done in each of those categories.

Common Forms of Process Evaluations

Description and assessment of program process are quite common in program evaluation, but the approaches used are varied, as is the terminology used. Such assessments may be conducted as one-shot endeavors or may be periodic so that information is produced regularly over an extended period of time, thus constituting program process monitoring. Process evaluations may be conducted by evaluators outside or inside the program organization or be set up as management tools with little involvement by professional evaluators. They may focus strictly on implementation fidelity to the program design or address broader questions of program coverage and the quality of services delivered. Moreover, their purpose may be to provide feedback for managerial purposes, to demonstrate accountability to sponsors and decision makers, to provide a freestanding process evaluation, or to augment an impact evaluation. Amid this variety, we further discuss the two principal forms of program process studies: individual process evaluations and continuous program monitoring.

Process Evaluation

Individual process evaluations are typically conducted by evaluation specialists as separate projects that will involve program personnel but are not integrated into their regular duties. When completed, and often while under way, process evaluation generally provides information about program performance to program managers and other stakeholders, but is not a regular and continuing part of a program's operation. [Exhibit 4-A](#) describes a process evaluation of a group of leadership academies designed to train principals to serve effectively in low-performing schools.

Exhibit 4-A Process Evaluation of Regional Leadership Academies

With federal funding, North Carolina established three Regional Leadership Academies (RLAs) to prepare principals to lead and reform low-performing schools throughout the state. Each academy was required to develop a plan describing how it would perform its major functions. The process evaluation focused on four questions:

1. Do RLAs recruit appropriate individuals to attend the academies relative to their intended target population?
2. Have the RLAs followed their plans for selective admission of program participants?
3. Is the training of school leaders in each RLA consistent with the program plan?
4. Do RLA graduates find placements in the intended leadership roles in low-performing schools and districts?

The evaluation team used three data sources for the process assessment: (a) administrative data from the state education agency, (b) semiannual surveys of program participants, and (c) observations of program activities, including weekly content seminars, advisory board meetings, mentor principal meetings, affiliated school districts' selection processes, induction support sessions, and specialized training opportunities.

The process evaluation found that the RLAs followed through on the activities specified in their plans with regard to recruitment, selective admission of participants, and provision of training that increased participants' rating of their own skills, and that graduates were being placed in lower performing schools. Specific findings included the following:

- The RLAs admitted 189 participants from a total of 962 applications, for an overall highly selective acceptance rate of less than 20%.
- The RLA participants were 71% female and 42% underrepresented minorities, representing greater diversity than the current population of principals in the state.
- The RLAs provided training on instructional leadership skills, resiliency skills, and school transformational skills using a curriculum that emphasized the challenges of working in high-need schools and the leadership strategies needed to turn around low performance in these schools.
- The participants, on average, gave positive ratings to their perceived gains in the competence and skills needed to lead reform in low-performing schools. Those ratings increased from midway between developing and proficient when they entered the RLAs to midway between proficient and accomplished after the 1st year.
- The participants served their yearlong internships in schools that averaged 66% economically disadvantaged students, and immediately following program completion 79% of the participants were employed as principals or assistant principals.

Source: Adapted from Brown, Stewart, and D'Amico (2014).

As an evaluation approach, process evaluation plays two major roles. First, it can stand alone as an evaluation of a program in circumstances in which the only questions at issue are about the integrity of program operations, service delivery, and other such matters. There are several kinds of situations that fit this description. A stand-alone process evaluation might be appropriate for a relatively new program, for instance, to answer questions about how well it has established its intended operations and services and to provide useful feedback to program managers and sponsors.

The process evaluation presented in [Exhibit 4-A](#) is an example of an evaluation of a new initiative, RLAs in their first 3 years of operation.

In the case of a more established program, a process evaluation might be initiated when questions arise about how well the program is organized, the quality of its services, or the success with which it is reaching the target population. A process evaluation may also constitute the major evaluation approach to a program charged with delivering a service known or presumed to be effective, so that the most significant performance issue is whether that service is being delivered properly. In a managed care environment, for instance, process evaluation may be used to assess whether prescribed medical treatment protocols are being followed for patients in different diagnostic categories.

The second major role of process evaluation is as a complement to an impact evaluation. Indeed, it is generally not advisable to conduct an impact evaluation without including at least a minimal process evaluation. Because maintaining an operational program and delivering appropriate services on an ongoing basis are formidable challenges, it is not generally wise to take adequate program implementation for granted. A full impact evaluation, therefore, often includes a process component to determine the quality and quantity of services the program provides so that that information can be integrated with findings from the impact of those services. In particular, impact evaluations are more informative when accompanied by an assessment of the fidelity of program implementation.

Implementation fidelity is the extent to which the program adheres to the program theory and design and usually includes such particulars as the amount of service received by the participants and the quality with which those services are delivered. Implementation fidelity information contributes to an impact evaluation in several ways. First, it helps establish that implementation was sufficient to plausibly produce the program effects that the impact evaluation will attempt to detect. Conversely, if implementation is poor, that fact offers a possible explanation if the expected program effects are not found. Second, the implementation data provide descriptive documentation of the nature of the program that does or does not produce the intended effects. Little sense can be made of impact

evaluation results without a clear picture of the nature of the program that produced those results. Third, the program effect estimates generated by most impact evaluation designs involve a comparison of outcomes for program participants with those for selected nonparticipants. The extent of the contrast in program experiences between those groups is thus a central issue in those designs. Implementation data characterize the program arm of that comparison and can often be adapted to determine the extent to which nonparticipants were exposed to services similar to those provided to program participants.

In [Exhibit 4-B](#), we list six components of a process evaluation that includes an assessment of implementation fidelity from a recent detailed book on the topic (Saunders, 2016, p. 148).

Exhibit 4-B Six Components of Comprehensive Process Evaluation

Saunders (2016) lists six components that should be considered when planning a process evaluation including implementation fidelity and the purpose each serves:

Plan Component	Purpose
Fidelity	Extent to which the program was implemented consistently with underlying theory, design, and philosophy
Dose delivered	Amount or number of intended units of each component delivered by the program staff
Dose received	Extent to which participants actively engage with, interact with, are receptive to, and/or use materials or resources; can include initial use and continued use
Satisfaction	Participant (can be direct and indirect participants) satisfaction with program and interactions with staff
Reach	Proportion of the priority intended beneficiaries who participate in the program; include documentation of barriers to participation
Recruitment	Procedures used to approach and attract participants at individual or organizational level; includes maintenance of participant involvement in program activities and measurement components of the evaluation

Source: Adapted from Saunders (2016).

Source: Adapted from Saunders (2016).

Process Monitoring and Administrative Data Systems

The second broad form of program process evaluation consists of continuous monitoring of indicators of selected aspects of program process. Such process monitoring can be a useful tool for supporting effective management of social programs by providing regular feedback about how well the program is performing its critical functions. This type of feedback allows managers to take corrective action when problems arise and can also provide stakeholders with regular updates about program performance. For these reasons, a form of process assessment is often integrated into routine **administrative data systems** so that appropriate data are obtained, compiled, and periodically summarized. In such cases, process evaluation captures information primarily from administrative data collected for intake, service documentation, and billing purposes. [Exhibit 4-C](#) provides an example in which electronic patient records are used to monitor medical practices for diabetes patient care throughout a network of providers.

Exhibit 4-C A Monitoring System for a Multifaceted Diabetes Intervention in an Integrated Delivery System

Diabetes is a chronic illness, affecting approximately 7% of the U.S. population, that requires coordinated medical care and patient self-management to decrease the risk for downstream complications. National guidelines for appropriate patient care exist, yet in practice actual care often fails to meet these guidelines. Monitoring physicians' compliance with patient care guidelines and providing feedback has been shown to be an effective strategy to improve physician adherence to those guidelines.

In a large network of physicians providing care for patients with diabetes, electronic patient records were compiled into an ongoing monitoring system that generated computerized reminders about diabetes practice guidelines and monthly reports on compliance with specific practices and a bundle of nine high-priority practices. Significant increases were seen in compliance with diabetes care guidelines.

- Vaccination for pneumococcal disease and influenza improved from 57% to 81% and from 55% to 71%, respectively.
- The percentage of patients with ideal glucose control increased from 32% to 35%, and blood pressure control improved from 40% to 44%.
- The overall number of patients receiving all nine high-priority practices and measurements within the desired range improved from 2.4% to 6.5%.

While careful to note that improved care is not sufficient to conclude that patients health also improved, the authors summarized the reaction to the care monitoring data by saying,

“It was distressing to our physicians that their ‘bundle score’ was initially low. We believe that this response created an early momentum for practice improvements. This low initial score also made it clear that increased physician vigilance and hard work alone would not result in success and encouraged team-based approaches to care.”

Source: Adapted from Weber, Bloom, Pierdon, and Wood (2008).

Administrative data systems routinely collect information on a client-by-client basis about services provided, staff providing the services, diagnosis or reasons for program participation, sociodemographic data, treatments and their costs, outcome status, and so on. Some systems bill clients (or funders), issue payments for services, and store other information, such as a client’s treatment history and current participation in other programs. Administrative data systems have become the major data source in many instances for process evaluation. Even when a program’s data system is not configured to completely fulfill the requirements of a thoroughgoing process evaluation, it may nonetheless provide a large portion of the information an evaluator needs for such purposes. Data retrieved from these systems are likely to be accurate when the data also serve administrative purposes, for example, when diagnostic information on clients is used for billing.

Perspectives on Program Process Monitoring

There is and should be considerable overlap in the purposes of process evaluation, whether it is driven by the information needs of evaluators, program managers, policymakers, sponsors, or stakeholders. Ideally, the assessment or monitoring activities undertaken should meet the information needs of all these groups. In practice, however, limitations on time and resources may require giving priority to one set of information needs over another. More generally, we can distinguish three perspectives on program process that vary in emphasis and overall purpose.

Process Assessment From the Evaluator's Perspective

A number of practical considerations underlie the need for evaluation researchers to assess program process. All too often a program's impact is diminished and, indeed, sometimes reduced to zero because the intervention was not delivered as designed, not delivered to the right target population, or both. There is good reason to believe that many failures of programs to produce the intended effects are due to implementation problems rather than to lack of potentially effective service concepts. As noted earlier, therefore, process evaluations are essential to understanding and interpreting impact findings. Knowing what took place is a prerequisite for explaining or hypothesizing why a program did or did not work as expected.

Process Assessment From an Accountability Perspective

Process assessment information is also critical for those who sponsor and fund programs. Program managers have a responsibility to inform their sponsors and funders of the activities undertaken, the degree of implementation of the programs as designed, problems encountered, and what the future holds (see [Exhibit 4-D](#) for one perspective on this matter). However, evaluators frequently are mandated to provide the same or similar information as an independent and objective respondent about what is going on in a particular program. This may be in the context of formative evaluation to guide program improvement, but it may also be for accountability purposes if program sponsors are concerned that program performance may not be strong enough to justify further funding or support.

Exhibit 4-D Describing Implementation of an Evidence-Based Intervention to Reduce Teen Pregnancy

When process monitoring is undertaken for accountability purposes, it is usually important to describe what was done by the program, who and how many were served, and details of the service delivery. This example involves a process evaluation that described the implementation of evidence-based interventions through multicomponent, community-wide initiatives to reduce teen pregnancy. Surveys from 2011 through 2014 were used to collect information about the capacity of state and community-based organizations to support implementation of these interventions, including documenting the characteristics of the interventions and information about the participants.

The survey results showed that over the period represented, the state and community-based organizations increased their capacities to support program partners in delivering evidence-based interventions. Those organizations provided 5,015 hours of technical assistance and training on topics including ensuring adequate capacity, process and outcome evaluation, program planning, and continuous quality improvement. Program partners increased the number of youth reached by an evidence-based intervention in the targeted communities from 4,304 in the 1st year of implementation in 2012 to 19,344 in 2014. In 2014, 59% of the youth received sexuality education programs, with smaller percentages receiving abstinence-based, youth development, and clinic-based programs. The majority of youth, 72%, were reached through schools and 16% through community-based organizations.

The authors concluded, “Building and monitoring the capacity of program partners to deliver [evidence-based interventions] through technical assistance and training is important. In addition, partnering with schools leads to reaching more youth.”

Source: Adapted from House, Tevendale, and Martinez-Garcia (2017).

Government sponsors and funders often operate in the glare of the news media and social media. Their actions are also visible to the legislative groups that authorize programs and to government watchdog organizations. For example, at the federal level, the Office of Management and Budget, part of the executive branch, wields considerable authority over program development, funding, and expenditures. The U.S. Government Accountability Office, an arm of Congress, advises members of the House and Senate on the utility of programs and in some cases conducts evaluations. Both state governments and those of large cities have analogous oversight groups. No social program that receives outside funding, whether public or private, can expect to avoid scrutiny and escape demands for **accountability**. Process evaluations make an important contribution in this context by helping identify programs that are performing well in providing the services for which they are responsible and those that are not performing well.

Process Assessment From a Management Perspective

Management-oriented process assessment is often concerned with the same questions as process assessment for accountability; the differences lie mainly in the applications of the findings. For accountability, process evaluation results are used primarily by decision makers, sponsors, and other stakeholders in oversight roles to judge the appropriateness of program activities and to consider whether a program should be continued, expanded, or contracted. In contrast, process evaluation results for which program managers are the main recipients are generally used for identifying and troubleshooting performance problems and taking corrective action. In that regard, their application is for purposes of sustaining good performance and improving performance where it is needed.

Process assessment from a management perspective is particularly vital during the implementation and pilot testing of new programs, especially innovative ones. No matter how well planned such programs may be, unexpected problems and shortcomings often surface early in the course of implementation. Program designers and managers need to know rapidly and fully about these problems so changes can be made to address them as soon as possible. Suppose, for example, that a medical clinic intended to help working mothers is open only during daylight hours. Monitoring may disclose that however great the demand for clinic services, the clinic's hours of operation effectively screen out most of the target population. Or suppose that a program is predicated on the assumption that severe psychological problems are prevalent among children who act out in school. If it is found early on that most such children do not in fact have serious disorders, the program can be modified accordingly.

For programs that have moved beyond the development stage to actual operation, program process assessments serve management needs by providing information on service delivery and coverage (the extent to which a program reaches its intended target population), and perhaps the reactions of participants to their experience with the program. Adjustments in the

program operation may be necessary when process information indicates, for example, that the intended beneficiaries are not being reached, that program costs are greater than expected, or that staff workloads are either too heavy or too light. This feedback is so useful to managers aiming to administer a high-performing program that it is desirable to receive it regularly rather than being limited to a single or only occasional process evaluation. Well-managed programs, therefore, often implement process monitoring systems that provide such performance data routinely, often integrated with a more general management information system.

Where process information is to be used for both managerial and evaluation purposes, some problems must be anticipated. How much information is sensible to collect and report, in what forms, at what frequency, with what reliability, and with what degree of confidentiality are among the issues on which evaluators and managers may disagree. For example, an experienced manager of a nonprofit children's recreational program may feel that the highest priority is weekly attendance information. The evaluator, however, may prefer to aggregate the attendance data monthly or even quarterly to smooth out uninformative short-term fluctuations.

Another concern is the matter of proprietary claims on the data. For the manager, performance data on, say, a novel program innovation should be kept confidential and shared only with the board of directors. The evaluator may believe that transparency is important to the integrity of the process evaluation and want to disseminate the results more broadly. Or a serious drop in clients from a particular ethnic group may result in the administrator of a program immediately replacing the director of professional services, whereas the evaluator's reaction may be to investigate further to try to determine why the drop occurred. As with all relations between program staff and evaluators, negotiation of such matters is essential. If the evaluator is not an employee of the agency, the administrators of the agency and evaluator will normally develop a memorandum of agreement that provides details on the purposes for which the data can be used, who has rights to use the data, and agreements about communicating findings drawn from the data. In addition to the memorandum of agreement, evaluators should also ensure that proper protection of human subjects is in place for any use of

administrative data for evaluation purposes. In [Chapter 11](#), we describe such memoranda and the human subjects review in more detail.

Note that there are many aspects of program management and administration (such as complying with tax regulations and employment laws or negotiating union contracts) that few evaluators have any special competence to assess. Proper expertise will need to be included on the evaluation team if such matters are within the scope of a process evaluation. More generally, capable process evaluation will almost always require subject matter expertise in the content area addressed by the program. The lead evaluator need not have that expertise, but someone on the evaluation team or consulting with the lead evaluator who does have that expertise should be involved in planning the process evaluation, reviewing the resulting data, and interpreting their implications for program performance.

In the remainder of this chapter, we concentrate on the concepts and methods pertinent to evaluating program process in the domains of service utilization and program organization. It is in this area that the competencies of trained evaluators are most relevant.

Assessing Service Utilization

A critical issue in program process evaluation is ascertaining the extent to which the intended target population actually receives program services. Managing a project effectively requires that participation of intended beneficiaries be sustained at an acceptable level and that corrective action be taken if it falls below that level. Assessing service utilization is particularly critical for interventions in which program participation is voluntary or participants must learn new procedures, change habits, or take instruction. For example, community mental health centers designed to provide a broad range of services often fail to attract a significant proportion of those who could benefit from their services. As shown in a classic evaluation study, even homeless patients recently discharged from psychiatric hospitals and encouraged to make use of the services of community mental health centers often failed to contact the centers (Rossi, Fisher, & Willis, 1986). Similarly, a program designed to provide information to prospective home buyers might find that few persons seek the services offered. Hence, program developers and managers need to be concerned with how best to engage and motivate members of the target population to seek out the program and participate in it. Depending on the particular situation, they might, for example, need to build outreach efforts into the program or pay special attention to the geographic placement of program sites.

Coverage and Bias

Service utilization issues typically break down into questions about coverage and bias. Whereas **coverage** refers to the extent to which participation by the target population achieves the levels specified in the program design, **bias** is the degree to which some subgroups participate in greater proportions than others. Clearly, coverage and bias are related. A program that reaches all the intended participants and no others is obviously not biased in its coverage. But because few social programs achieve such total coverage, bias is a common concern.

Bias can arise from self-selection; that is, some subgroups may voluntarily participate more frequently than others. It can also derive from program actions. For instance, program personnel may react favorably to some clients while discouraging others. One temptation commonly faced by programs is to select the most success prone targets, with the expectation, therefore, of getting positive outcomes that make the program look good. Known as *creaming*, this situation frequently occurs because of the self-interests of one or more stakeholders (an example is described in [Exhibit 4-E](#)). Finally, bias may result from such unforeseen influences as the location of a program office or the hours during which it operates such that some subgroups have more convenient access than others.

Although there are many social programs, such as the federal food stamp program, that aspire to serve all or a very large proportion of a defined target population, typically programs do not have the resources to provide services to more than a fraction of potential beneficiaries. Program staff and sponsors can correct this problem by defining the characteristics of the target population more sharply and by using resources more effectively. For example, establishing a health center to provide medical services to persons in a defined community who do not have regular sources of care may result in such an overwhelming demand that many of those who want services cannot be accommodated. The solution might be to add eligibility criteria that weight such factors as severity of the health problem, family size, age, and income to reduce the size of the target population to manageable proportions while still serving persons with the greatest need. In some

programs, such as the Special Supplemental Nutrition Program for Women, Infants, and Children or housing vouchers for the poor, undercoverage is a systemic problem; Congress has never provided sufficient funding to cover all who are eligible.

Exhibit 4-E Charter School Creaming of Students

Charter schools are publicly funded but operate outside of the traditional public school system. In contrast to standard public schools, which serve the school-aged children in their neighborhoods, parents and students must choose to attend charter schools, and the students and their families must meet the schools' requirements in order to enroll. Critics of charter schools charge that they take resources away from public schools and that they may implement practices that exclude some children from admission or push out those who are more difficult to teach.

In a study by the RAND Corporation, evaluators assessed creaming by charter schools using administrative data from seven different states and municipalities. In terms of prior test scores, the students transferring into charter schools were near or below local averages in every geographic location included in the study. Although the students transferring into the charter schools were predominately African American at most sites, the racial composition of the charter schools was similar to that of the local public schools from which the students came. However, the study found some evidence that African American students transferring to charter schools in most locations moved to schools with higher concentrations of African American students than in the schools from which they transferred.

Another study led by the same evaluator analyzed administrative data from a large municipality to see if lower performing students were more likely to transfer out of charter schools than higher performing students. That study found no evidence of that pattern among students leaving charter schools. The evaluators went further to investigate the transfer patterns for low-performing students in each school in the district, reporting that "we found only 15 out of more than 300 schools district-wide in which below-average students were more likely to transfer out than above average students at rates of 10 percent or more. Of these, only one is a charter school, and that school focuses on students at-risk of dropping out."

These two studies thus did not support the claim that charter schools were pushing out low-performing students or creaming higher performing students relative to the public noncharter schools.

Sources: Adapted from Zimmer and Guarino (2013) and Zimmer et al. (2009).

The opposite effect, overcoverage, also occurs. For instance, the TV program *Sesame Street* has consistently captured audiences far exceeding the intended targets (economically disadvantaged preschoolers), including children who are not at all disadvantaged and even adults. Because these

additional audiences are reached at no additional cost, this overcoverage is not a financial drain. It may, however, thwart one of *Sesame Street's* original goals, which was to lessen the gap in learning between economically disadvantaged children and their more advantaged peers.

The most common coverage problem in social programs, however, is the failure to achieve high target population participation either because of bias in the way targets are recruited or retained or because potential clients are unaware of the program, are unable to use it, or reject it. For example, in most employment training programs only small minorities of those eligible by reason of unemployment ever attempt to participate, and certain subpopulations of those eligible may have dramatically low rates relative to other eligible subgroups. In [Exhibit 4-F](#), the relatively low coverage rates of individuals with disabilities in employment programs and their overrepresentation in safety net programs are assessed. Similar situations occur in mental health, substance abuse, and numerous other programs. We turn now to the question of how program coverage and bias might be measured as part of a program process evaluation.

Exhibit 4-F The Coverage of Federal Safety Net and Employment Programs for Individuals With Disabilities

With many federal programs facing budget shortages, this study assessed the coverage of safety net and employment programs, with a focus on participation by individuals with disabilities. The 2009 Current Population Survey–Annual Social and Economic Supplement, conducted by the Census Bureau, allowed researchers to identify households with persons with and without disabilities and determine program participation rates on the basis of self-reports. Focusing on the working-age population, individuals between 24 and 61, the study revealed that people with disabilities represented one third of the persons who participated in safety net programs, with 65% of individuals with disabilities participating in one or more of those programs. This is comparable with a 17% participation rate of persons without disabilities. The results also showed that only 3% of low-income, nonworking, safety net participants with disabilities used employment services, which compares with 8% of low-income, nonworking, safety net participants without disabilities. The authors suggest that increasing coordination of employment services for individuals with disabilities so as to obtain greater coverage of that subgroup might improve their well-being and potentially reduce the financial strain on safety net programs.

Source: Based on Houtenville and Brucker (2014).

Measuring Coverage

Program managers and sponsors alike need to be concerned with both undercoverage and overcoverage. Undercoverage is measured by the proportion of the individuals eligible for a program who actually participate in it. Overcoverage is often expressed as the number of program participants who are not in need compared with the total number of participants in the program. Efficient use of program resources requires both maximizing the number served who are in need and minimizing the number served who are not in need.

The problem in measuring coverage is almost always the inability to specify the number in need, that is, the size of the target population. The needs assessment procedures described in [Chapter 2](#), if carried out as an integral part of program planning, usually minimize this problem. In general there are three sources of information that can be used to assess the extent to which a program is serving the appropriate target population: program records, surveys of program participants, and community surveys.

Program Records

Almost all programs keep records on the individuals served. Data from well-maintained administrative record systems can often be used to estimate program bias or overcoverage. For instance, information on the various screening criteria for program intake may be tabulated to determine whether the units served are the ones specified in the program's design. Suppose the targets of a family planning program are women less than 50 years of age who have been residents of the community for at least 6 months and who have two or more children under age 10. Records of program participants can be examined to see whether the women actually served are within the eligibility limits and the degree to which particular age or parity groups are under- or overrepresented. Such an analysis might also disclose bias in program participation in terms of the eligibility characteristics or combinations of them.

However, even in this digital age programs differ widely in the quality and extensiveness of their records and in the sophistication involved in storing and maintaining them. Moreover, the feasibility of maintaining complete, ongoing record systems for all program participants varies with the nature of the intervention and the available resources. In the case of medical and mental health systems, for example, sophisticated electronic record systems have been developed for managed care purposes that would be impractical for many other types of programs.

In measuring target population participation, the main concerns are that the data are accurate and reliable. It should be noted that all record systems are subject to some degree of error. Some records will contain incorrect or outdated information, and others will be incomplete. The extent to which unreliable records can be used for decision making depends on the kind and degree of their unreliability and the nature of the decisions in question. Clearly, critical decisions involving significant outcomes require better records than do less weighty decisions. Whereas a decision on whether to continue a project should not be made on the basis of data derived from partly unreliable records, data from the same records may suffice for a decision to change an administrative procedure. One overarching principle to invoke when considering the use of administrative records is that they are likely to be most accurate when the data elements of interest for the evaluation are used for program administrative purposes. For example, an evaluator may use records of teachers' salary payouts to measure teacher turnover. If the records are used in disbursing monthly paychecks and these payments are audited, they are likely to be highly accurate about dates of employment.

If program records are to serve an important role in evaluation of program processes, it is usually prudent to examine the records for accuracy before using them as a data source. For example, records might be sampled to determine whether each program participant has a single record, whether the data on each record are complete, and whether rules for completing them have been followed.

Surveys

An alternative to using program records to assess target population participation is to conduct surveys of program participants. Sample surveys may be desirable when the required data cannot be obtained as a routine part of program activities or when the size of the population group is large and it is more economical and efficient to undertake a sample survey than to obtain data on the entire population.

For example, a special tutoring project conducted primarily by parents may be set up in only a few schools in a community. Children in all schools may be referred, but the project staff may not have the time or the training to administer appropriate educational skills tests and other such instruments that would document the characteristics of the children referred and enrolled. Lacking such complete records, an evaluator could administer tests to a sample of the children receiving tutoring to estimate the appropriateness of the selection procedures and assess whether the project is serving the designated target population.

When projects are not limited to selected, narrowly defined groups of individuals but instead take in entire communities, the most efficient and sometimes the only way to examine whether the presumed population at need is being reached is to conduct a community survey. Various types of health, educational, recreational, and other human service programs are often community-wide, although their intended target populations may be selected groups, such as delinquent youths, the aged, or women of childbearing age. In such cases, surveys are the major means of assessing whether targets have been reached.

The evaluation of the *Feeling Good* television program years ago illustrates the use of surveys to provide data on a project with a national audience. The program, an experimental production of the Children's Television Workshop (the producer of *Sesame Street*), was designed to motivate adults to engage in preventive health practices. Although it was accessible to homes of all income levels, its primary purpose was to motivate low-income families to improve their health practices. The Gallup organization conducted four national surveys, each of approximately 1,500 adults, at different times during the weeks *Feeling Good* was televised. The data provided estimates of the size of the viewing audiences and of the viewers'

demographic, socioeconomic, and attitudinal characteristics (Mielke & Swinehart, 1976). The major finding was that the program largely failed to reach the target group, and the program was discontinued.

To measure coverage of U.S. Department of Labor programs, such as training and public employment, the department started a periodic national sample survey. The Survey of Income and Program Participation is now carried out by the Census Bureau and measures participation in social programs conducted by many federal departments. This large survey, now a 3-year panel covering 21,000 households, ascertains through personal interviews whether each adult member of the sampled households has ever participated or is currently participating in any of a number of federal programs. By contrasting program participants with nonparticipants, the survey provides information on the programs' biases in coverage. In addition, it generates information on the uncovered but eligible members of the target populations.

Assessing Bias: Program Users, Eligibles, and Dropouts

An assessment of bias in program participation can be undertaken by examining differences between individuals who participate in a program and either those who drop out or those who are eligible but do not participate at all. In part, the drop-out rate from a program may be an indicator of dissatisfaction with the program. It also may indicate conditions in the community that militate against full participation. For example, in certain areas lack of adequate transportation may prevent those who are otherwise willing and eligible from participating in a program.

It is important to be able to identify the particular subgroups within the target population who either do not participate at all or do not follow through to full participation. Such information not only is valuable in judging the worth of the effort but also is needed to develop hypotheses about how a program can be modified to attract and retain a larger proportion of the target population. Thus, the qualitative aspects of participation may be important not only for process evaluation purposes but also for subsequent program planning.

Data about dropouts may come either from administrative records or from surveys designed to identify nonparticipants. However, community surveys usually are the only feasible means of identifying eligible persons who have not participated in a program. The exception, of course, is when adequate information is available about the entire eligible population prior to the implementation of a program (as in the case of data from a census or screening interview).

In [Chapter 10](#), we describe methods of analyzing the costs and benefits of programs to arrive at measures of economic efficiency. Clearly, for calculating costs it is important to have estimates of the size of populations at need or risk, the groups who start a program but drop out, and the ones who participate to completion. The same data may also be used in estimating benefits. In addition, such data are useful in judging whether a

program should be continued and whether it should be expanded. Furthermore, project staff require this kind of information to meet their managerial and accountability responsibilities. Although data on program participation cannot substitute for knowledge of impact in judging either the efficiency or the effectiveness of projects, an adequate description of the extent of participation by the target population is relevant for interpreting the estimates of impact.

Assessing Organizational Functions

Monitoring of the critical organizational functions and activities of a program focuses on whether the program is performing well in managing its efforts and using its resources to accomplish its essential tasks. Chief among those tasks, of course, is delivering the intended services to the target population. In addition, programs have various support functions that must be carried out to maintain the viability and effectiveness of the organization, for example, fund-raising, promotion and advocacy, and governance and management. Program process monitoring seeks to determine whether a program's actual activities and arrangements sufficiently approximate the intended ones.

Once again, program process theory as described in [Chapter 3](#) is a useful tool in designing a process assessment. In this instance, what was called the organizational plan in that chapter is the relevant component. A fully articulated process theory will identify the major program functions, activities, and outputs and show how they are related to one another and to the organizational structures, staffing patterns, and resources of the program. This depiction provides a map to guide the evaluator in identifying the significant program functions and the preconditions for accomplishing them. Program process evaluation then becomes a matter of identifying and measuring those activities and conditions most essential to a program's ability to carry out its duties.

The Delivery System

A program's delivery system can be thought of as a combination of pathways and actions undertaken to provide an intervention. It usually consists of a number of separate functions and relationships. As a general rule, it is wise to assess all the elements unless previous experience with certain aspects of the delivery system makes that unnecessary. Two concepts are especially useful for evaluating the performance of a program's delivery system: specification of services and accessibility.

Specification of Services

A specification of services is desirable for both planning and assessment purposes. This consists of specifying the actual services provided by the program in operational (measurable) terms. The first task is to define each kind of service in terms of the activities that take place and the providers who participate. When possible, it is best to separate the various aspects of a program into separate, distinct services. For example, if a program providing technical education for school dropouts includes literacy training, carpentry skills, and a period of on-the-job apprenticeship work, it is advisable to separate these into three services for evaluation purposes. Moreover, for estimating program costs in cost-benefit analyses and for fiscal accountability, it is often important to attach monetary values to different services. This step is important when the costs of several programs will be compared or when the programs receive reimbursement on the basis of the number of units of different services that are provided.

For program process evaluation, simple, specific services are easier to identify, count, and record. However, complex elements often are required to design an implementation that is consistent with a program's objectives. For example, a clinic for children may require a physical exam on admission, but the scope of the exam and the tests ordered may depend on the characteristics of each child. Thus, the item "exam" is a service, but its components cannot be broken out further without creating a different definition of the service for each child examined. The strategic question is

how to strike a balance, defining services so that distinct activities can be identified and counted reliably while, at the same time, the distinctions are meaningful in terms of the program's objectives.

In situations in which the nature of the intervention allows a wide range of actions that might be performed, it may be possible to describe services primarily in terms of the general characteristics of the service providers and the time they spend in service activities. For example, if a program places master craftspeople in a low-income community to instruct community members in ways to improve their dwelling units, the craftspeople's specific activities will vary greatly from one household to another. They may advise one family on how to frame windows and another on how to shore up the foundation of a house. Any process assessment attempting to document such services could describe the service activities only in general terms and by means of examples. It is possible, however, to specify the characteristics of the providers—for example, that they should have 5 years of experience in home construction and repair and knowledge of carpentry, electrical wiring, foundations, and exterior construction—and the amount of time they spend with each service recipient.

Indeed, services are often defined in terms of units of time, costs, procedures, or products. In a vocational training project, service units may refer to hours of counseling time provided; in a program to foster housing improvement, they may be defined in terms of amounts of building materials provided; in a cottage industry project, service units may refer to activities, such as training sessions on how to operate sewing machines; and in an educational program, the units may be instances of the use of specific curricular materials in classrooms. All these examples require an explicit definition of what constitutes a service and, for that service, what units are appropriate for describing the amount of service.

Accessibility

Accessibility is the extent to which structural and organizational arrangements facilitate participation in a program. All programs have strategies of some sort for providing services to the appropriate target populations. In some instances, being accessible may simply mean opening

an office and operating under the assumption that the designated target population will appear and make use of the services provided at the site. In other instances, however, ensuring accessibility requires outreach campaigns to recruit participants, transportation to bring persons to the intervention site, and efforts during the intervention to minimize dropouts. For example, in many large cities, special teams are sent out into the streets on very cold nights to persuade homeless persons sleeping in exposed places to spend the night in shelters. In [Exhibit 4-G](#), we describe the evaluation of an innovative pilot program to curb summer learning loss by providing children in low-income communities with access to books. The books were distributed through vending machines free of charge, with important process evaluation questions about children retrieving the books and subsequently reading them.

A number of process evaluation questions arise in connection with accessibility, some of which relate only to the delivery of services and some of which have parallels to the previously discussed topic of service utilization. The primary issue is whether program actions are consistent with the design and intent of the program with regard to facilitating access. For example, is a Spanish-speaking staff member always available in a mental health center located in an area with a large Hispanic population?

Also, are potential participants matched with the appropriate services? It has been observed, for example, that community members who initially make use of emergency medical care services for appropriate purposes may subsequently use them for general medical care. Such misuse of emergency services may be costly and reduce their availability to other community members. A related issue is whether the access strategy encourages differential use by participants from certain social, cultural, and ethnic groups, or whether there is equal access for all potential participants.

Program Support Functions

Although providing the intended services is presumed to be a program's main organizational function, and one essential to assess, most programs also perform important support functions that are critical to their ability to maintain themselves and continue to provide service. These functions are of interest to program administrators, of course, but often they are also relevant to assessment by evaluators or outside decision makers. Vital support functions may include such activities as fund-raising; public relations to enhance the program's image with potential sponsors, decision makers, or the general public; staff training, including the training of the direct service staff; recruiting and retention of key personnel; developing and maintaining relationships with affiliated programs, referral sources, and other external collaborators; obtaining materials required for services; and general advocacy on behalf of the target population served.

Program process evaluation schemes can, and often should, incorporate indicators of vital program support functions along with indicators relating to service activities. In form, such indicators and the process for identifying them are no different than for program services. The critical activities first must be identified and described in specific, concrete terms resembling service units; for example, units of fund-raising activity and dollars raised, number, length, and quality of training sessions, number and characteristics of attendees at advocacy events, and the like. Measures are then developed that are capable of differentiating good from poor performance. These measures can then be included in the process evaluation or program monitoring procedures along with those dealing with other aspects of program performance.

Exhibit 4-G Summertime Distribution of Books for Children in Low-Income Communities

Noting persistent achievement gaps between economically disadvantaged children and their more affluent peers and the academic slide that occurs for lower performing children during the summer, a pilot book distribution program was established in four low-income neighborhoods. During the summer in both Detroit and Washington, D.C., age-appropriate books were placed in vending machines designed to dispense the books (see picture) at no cost. The vending machines were placed in high-traffic places near

churches or childcare centers and available to passers-by. Books were restocked frequently, and new titles, including fiction and nonfiction offerings, were added throughout the summer. Childcare centers and parents were notified of the availability of the books and the location of the machines.

The evaluators made a total of 48 two-hour observations of the activity around the vending machines and conducted short interviews with individuals who either retrieved books or viewed them without taking one. They also administered several short assessments, including book title recognition and pre- and postsummer assessments of children's reading skills.



During the summer, the vending machines distributed 64,435 books in total, 59% of which went to return users. On average, 180 people passed the sites over the 2-hour observation periods, and about 50 of them visited the vending machines. The visitors were primarily people of color, and the majority at each site were female. The percentage of repeat visitors ranged from 33% to 52%. The numbers of books obtained by children of different age ranges were similar, with slightly fewer for 10- to 14-year-olds. More than two thirds of the books distributed were fiction. Interestingly, children who visited the vending machines with adults were more likely to take a book and recognized more of the book titles from a list of titles.

In their conclusion, the study authors stated, "As our interviews revealed, the close proximity of books to where people were likely to traffic clearly had its benefits to many in these communities. Almost half of the people accessing books were repeat users. Many regarded these resources as a welcome contribution to the local neighborhood, and a necessary support to help spark their children's interest and skill in reading. At the same time, traffic patterns indicated that there were a substantial number of people who chose not to access books (40%). Their primary reason, according to our interviews, was a lack of interest in reading."

Source: Adapted from Neuman and Knapczyk (2018).

Summary

- Process evaluation is a form of evaluation designed to describe how a program is operating and to assess how well it performs its intended functions. It builds on program process theory, which identifies the critical components, functions, and relationships assumed necessary for the program to be effective.
- The criteria for assessing program process performance may include stipulations from the program theory, administrative standards, applicable legal, ethical, or professional standards, and after-the-fact judgment calls.
- Process evaluation may be conducted as a separate stand-alone project by evaluation specialists. It may also be an ongoing function involving repeated measurements over time—referred to as program process monitoring—that would typically be part of a program's management information system.
- A process evaluation is often carried out in conjunction with an impact evaluation to describe the program services presumably responsible for whatever effects the impact evaluation finds on the intended outcomes. In that context, the focus is typically on assessing the fidelity of implementation, that is, the extent to which the intended services are actually delivered and their amount and quality.
- Program process evaluation takes somewhat different forms and serves different purposes when undertaken from the perspectives of evaluation, accountability, and program management, but the types of data required and the data collection procedures used generally are similar. In particular, program process evaluation generally involves one or both of two domains of program performance: service utilization and organizational functions.
- Service utilization issues typically break down into questions about coverage and bias. The sources of data useful for assessing coverage are program records, surveys of program participants, and community surveys. Bias in program coverage can be revealed through comparisons of program participants from different subgroups and examination of the characteristics of eligible nonparticipants and program dropouts.
- Assessment of a program's organizational functions focuses on how well the program is organizing its efforts and using its resources to accomplish its essential tasks. Particular attention is given to identifying shortcomings in program implementation that prevent a program from delivering the intended services to the target population. Monitoring of organizational functions also includes attention to the delivery system and program support functions.

Key Concepts

Accessibility 110
Accountability 102
Administrative data system 98
Administrative standards 95
Bias 104
Coverage 104
Implementation fidelity 98
Monitoring and evaluation 92
Outcome monitoring 93
Process evaluation 92
Process monitoring 92

Critical Thinking/Discussion Questions

1. Explain what a process evaluation is. Describe the different areas of focus a process evaluation can have. What are some of the main reasons for undertaking a process evaluation?
2. Describe the common forms of process evaluations. How are they similar, and what are the major differences?
3. Define coverage and bias and explain how they are related and how they can be examined in a process evaluation.

Application Exercises

1. We provide a list of questions a process evaluation can be designed to answer. Choose a local social program and determine what information you would need to answer these questions. Include information such as the populations you would involve in your study and what methods you would use to collect the data.
2. Using the same local program, design a process evaluation using the six components listed in the text (fidelity, dose delivered, dose received, satisfaction, reach, and recruitment). How would you address each of the six components in your evaluation?

Chapter 5 Measuring and Monitoring Program Outcomes

Program Outcomes

Outcome Level, Outcome Change, and Program Effect

Identifying Relevant Outcomes

Stakeholder Perspectives

Program Impact Theory

Prior Research

Unintended Effects

Measuring Program Outcomes

Measurement Procedures and Properties

Reliability

Validity

Sensitivity

Choice of Outcome Measures

Monitoring Program Outcomes

Indicators for Outcome Monitoring

Pitfalls in Outcome Monitoring

Interpreting Outcome Data

Summary

Key Concepts

The previous chapter discussed how a program's process and operational performance can be monitored and assessed. The ultimate goal of all programs, however, is not merely to function well, but to bring about change—to affect some problem or social condition in beneficial ways. A program's objectives for change are characterized as outcomes by both the program and evaluators assessing program effects.

The outcomes a program aspires to influence are identified in the program's impact theory and reflect the goals and objectives stakeholders have for the program. Sensitive and valid measurement of those outcomes can be technically challenging but is essential to assessing a program's success. Once developed, outcome measures can also be used in ongoing outcome monitoring schemes to provide informative feedback to program managers. Interpreting the results of outcome measurement and monitoring, however, presents challenges to stakeholders and evaluators because most outcomes can be influenced by many factors other than the intervention provided by the program. This

chapter describes how program outcomes can be identified, measured, and monitored, and how the results can be properly interpreted.

Assessing a program's effects on the clients it serves and the social conditions it aims to improve is the most critical evaluation task because it deals with the bottom-line issue for social programs. No matter how well a program diagnoses the needs it aims to ameliorate, embodies a good theory of action, reaches its target population, and delivers apparently appropriate services, it cannot be judged successful unless it actually brings about some degree of beneficial change in the outcomes it addresses. Measuring those outcomes, therefore, is not only a core evaluation function but also a high-stakes activity for a program. For these reasons, it is a function evaluators must accomplish with great care to ensure that evaluation findings about program outcomes are valid and properly interpreted. For these same reasons, it is one of the more difficult and, often, politically charged tasks the evaluator undertakes.

Measuring and monitoring outcomes rarely constitute a stand-alone evaluation. In many cases, outcomes are included when evaluating or monitoring program process, as discussed in [Chapter 4](#). Such monitoring schemes for both program process and outcomes are often incorporated into management information systems that can help administrators guide effective program performance. With the onset of the digital age, the key performance indicators from these schemes are increasingly being depicted and periodically updated in data displays called *data dashboards* and made publicly available via the Internet. Measuring outcomes is also a key component of all impact evaluations. Beginning in this chapter and continuing through [Chapter 8](#), we consider how to identify the outcomes a program should be expected to change, how to devise measures of those outcomes that respond to change, and how to determine the program's impact on those outcomes. Consideration of these matters begins with the concept of a program *outcome*, so we first discuss this pivotal concept.

Program Outcomes

An **outcome** is the state of the target population or the social conditions with which a program intervenes on a characteristic or behavior the program might potentially affect. For example, the prevalence of smoking among teenagers is an outcome for an antismoking campaign in their high school, as are attitudes toward smoking among those who have not yet started to smoke. Similarly, school readiness might be an outcome for a preschool program, as would the body weight of people in the target population for a weight-loss program, the management skills of business personnel for a management training program, and the amount of pollutants in the local river for a crackdown by the local environmental protection agency.

Notice two things about these examples. First, outcomes are observable characteristics of the target population or social conditions, not of the program, and the definition of an outcome makes no direct reference to program actions. The services provided by a program or received by participants are often described as program “outputs,” which are not to be confused with *outcomes* as defined here. Thus, “receiving supportive family therapy” is not a program outcome but, rather, the receipt of a program service. Similarly, providing meals to housebound elderly persons is not a program outcome; it is service delivered. The nutritional quality of the meals consumed by the elderly and the extent to which they are malnourished, on the other hand, are outcomes in the context of a program that serves meals to that population. Put another way, outcomes always refer to characteristics that, in principle, could be observed for individuals or social conditions that have not received program services. We could assess the prevalence of smoking, school readiness, body weight, management skills, and water pollution for the respective situations even when there was no program intervention.

Second, the concept of an outcome does not necessarily mean that there has been any actual change on that outcome, or that any change that has occurred was caused by the program rather than some other influence. The prevalence of smoking among high school students may or may not have

changed since the antismoking campaign began, and the participants in the weight-loss program may or may not have lost weight. Furthermore, whatever changes did occur may have resulted from something other than the influence of the program. Perhaps the weight-loss program ran during a holiday season when people were prone to overeating. Or perhaps the teenagers decreased their smoking in reaction to news of the smoking-related death of a popular rock musician.

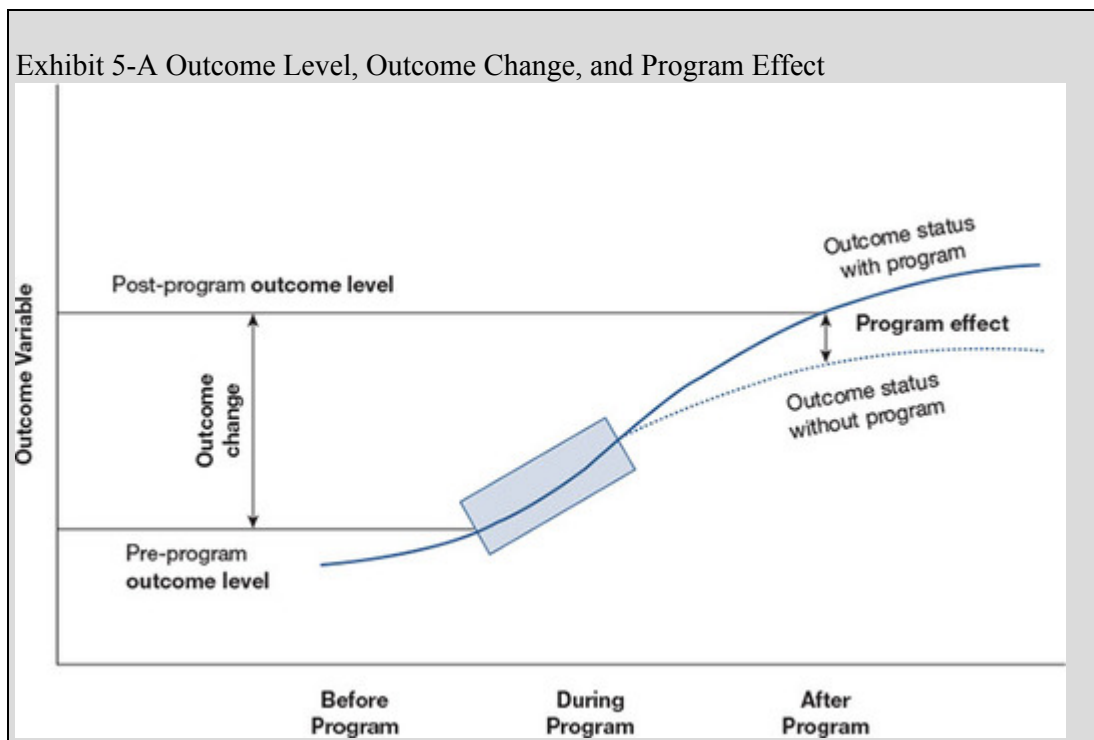
Outcome Level, Outcome Change, and Program Effect

These considerations lead to important distinctions in the use of the term *outcome*:

- **Outcome level** is the status of an outcome at some point in time (e.g., the prevalence of smoking among teenagers).
- **Outcome change** is the difference between outcome levels from one point in time to another (e.g., increase in the amount of smoking from the beginning to the end of the school year).
- **Program effect** is the difference between the outcome level for those exposed to the program and the outcome level they would have had if they had not been exposed to the program. It is the change on the outcome experienced by program participants that can be attributed directly and uniquely to the effects of the program as opposed to the influence of other factors.

Consider the graph in [Exhibit 5-A](#), which plots the values of an *outcome variable* on the vertical axis. An outcome variable is the set of values generated by measuring an outcome for a defined group of individuals or other units. It might, then, be the number of cigarettes each student in a high school reports smoking in the past month, or particulate matter per milliliter found in water samples drawn from the local river. The horizontal axis represents time, specifically, a period ranging from before any program exposure by those whose outcomes are measured until sometime afterward. The solid line in the graph shows the average outcome level for members of the target population who were exposed to the program. Note that change in the outcome over time is not depicted as a straight horizontal line but, rather, as a curved line that wanders upward over time. This is to indicate that smoking, school readiness, management skills, and other such outcomes are not expected to stay constant; they change as a result of natural causes and circumstances quite extraneous to the program. Smoking, for instance, tends to increase from the preteen to the teenage

years. Water pollution levels may fluctuate according to industrial activity in the region and weather conditions, and so forth.



At any point during the interval charted, the average value on the outcome variable for the individuals represented can be identified, indicating how high or low the group is with respect to that variable. This tells us the *outcome level*, often simply called the outcome, at a particular time. When measured after program exposure, it tells us something about how those individuals are doing: how many teenagers are smoking, the average level of school readiness among the preschool children, how much pollution is in the water, and so on. If all the teenagers are smoking after program exposure, we may be disappointed, and, conversely, if none are smoking, we may be pleased. All by themselves, however, these outcome levels do not tell us much about how effective the program was, though they may constrain the possibilities. If all the teens are smoking, for instance, we can be fairly sure the antismoking program was not a great success and possibly had adverse effects. If none of the teenagers are smoking, it is a strong hint that the program has worked, because we would not expect all of them to spontaneously stop on their own. Of course, such extreme outcomes are rarely found, and in most cases outcome levels alone cannot be interpreted with any confidence as indicators of a program's success or failure.

If we measure outcomes on the program recipients before and after their participation in the program, we can describe more than the outcome level—we can also discern outcome *change*. If the graph in [Exhibit 5-A](#) plots the school readiness of children in a preschool program, it shows less readiness before participation in the program and greater readiness afterward, a positive change. Even if school readiness after the program was not as high as the preschool teachers hoped, the direction of before-to-after change shows improvement. Of course, from this information alone we do not know what caused that change or whether the preschool program had anything to do with it. Preschool-aged children are in a developmental period when their cognitive and motor skills increase rapidly and naturally with or without a preschool program. Other factors may also be at work; for example, their parents may be reading to them and otherwise supporting their intellectual development and preparing them to enter school.

The dashed line in [Exhibit 5-A](#) shows the trajectory on the outcome variable that would have been observed if the program participants had not received the program. For the preschool children, for example, the dashed line shows how their school readiness would have increased without any exposure to the preschool program. The solid line shows how school readiness developed when they were in the program. A comparison of the two outcome lines indicates that school readiness would have improved even without exposure to the program, but not quite as much.

The difference between the outcome level attained with participation in the program and that which the same individuals would have attained had they not participated is the *program effect*, or the increment in the outcome that the program produced, also referred to as the program **impact**. This is the value added or net gain part of the outcome that would not have occurred without the program. It is the only part of the change on the outcome for which the program can rightfully take credit.

Estimation of program effects, or impact evaluation, is the most demanding form of evaluation. With the program effect defined as the difference between the outcome that occurred with program exposure and the outcome that would have occurred without program exposure, as illustrated in [Exhibit 5-A](#), it refers to outcomes for the same people (or other entities)

under mutually exclusive conditions. It is impossible for the same individuals to both participate and not participate in a program at the same time, and it follows that it is also impossible to observe both of the corresponding outcomes. To identify program effects, the evaluator must, therefore, measure the outcome after program participation and then somehow estimate what that outcome would have been without the program. The latter outcome must be estimated rather than measured because it is hypothetical for individuals who did, in fact, participate in the program. Developing valid inferences under these circumstances can be challenging. [Chapters 6, 7, and 8](#) describe the methodological tools and research designs evaluators have available for this daunting task.

Although assessment of outcome levels and outcome changes has rather limited utility for estimating program effects, the results are of some value to managers and stakeholder for monitoring program performance. This application of outcome measures will be discussed later in the chapter. For now we will continue our exploration of the concept of an outcome by discussing how they can be identified, defined, and measured for the purposes of evaluation.

Identifying Relevant Outcomes

The first step in developing measures of program outcomes is to identify very specifically what outcomes are relevant candidates for measurement. To do this, the evaluator must consider the perspectives of stakeholders about pertinent outcomes, the outcomes that are specified in the program's impact theory, and applicable prior research. The evaluator will also need to consider the possibility that there will be outcomes on which the program may produce unintended effects.

Stakeholder Perspectives

Various program stakeholders will have their own understandings of what the program is supposed to accomplish and, correspondingly, what outcomes they expect it to affect. The most direct sources of information about these outcomes usually are the stated objectives, goals, and mission of the program. Funding proposals and grants or contracts for services from outside sponsors also often identify outcomes the program is expected to influence. A common difficulty with information from these sources is a lack of the specificity and concreteness necessary to clearly identify and define the outcome. It thus often falls to the evaluator to translate input from stakeholders into workable form and negotiate with those stakeholders to ensure that the resulting outcome measures capture their expectations.

For the evaluator's purposes, an outcome description must indicate the pertinent characteristic, behavior, or condition the program is expected to change. However, as we discuss shortly, further specification and differentiation may be required as the evaluator moves from this description to selecting or developing measures of the outcome. [Exhibit 5-B](#) presents examples of outcome descriptions specific enough to be relatively serviceable for evaluation purposes.

Exhibit 5-B Examples of Outcomes Described Specifically Enough to Be Measured

- Juvenile delinquency: Behavior of youth under the age of 18 that constitutes chargeable offenses under applicable laws irrespective of whether the offenses are detected by authorities or the youth is apprehended for the offense
- Contact with antisocial peers: Friendly interactions while spending time with one or more youth of about the same age who regularly engage in behavior that is illegal and/or harmful to others
- Constructive use of leisure time: Engaging in behavior that has educational, social, or personal value during discretionary time outside of school and work
- Water quality: The absence of substances in the water that are harmful to people and other living organisms that drink the water or have contact with it
- Toxic waste discharge: The release of substances known to be harmful into the environment from an industrial facility in a manner that is likely to expose people and other living organisms to those substances
- Cognitive ability: Performance on tasks that involve thinking, problem solving, information processing, language, mental imagery, memory, and overall intelligence
- School readiness: Children's ability to learn at the time they enter school; specifically, the health and physical development, social and emotional

development, language and communication skills, and cognitive skills and general knowledge that enable a child to benefit from participation in formal schooling

- Positive attitudes toward school: The extent to which a child likes school, has positive feelings about attending, and is willing to participate in school activities

Program Impact Theory

A full articulation of the program impact theory, as described in [Chapter 3](#), is especially useful for identifying and organizing program outcomes. An impact theory, recall, expresses the outcomes of a social program as part of a logic model that connects the program's activities to proximal outcomes that, in turn, are expected to lead to other, more distal outcomes. If correctly described, this series of linked relationships among outcomes represents the program's assumptions about the critical steps between program services and the ultimate social benefits the program is intended to produce. It is thus especially important for the evaluator to draw on this portion of the program theory when identifying the outcomes that should be considered for measurement.

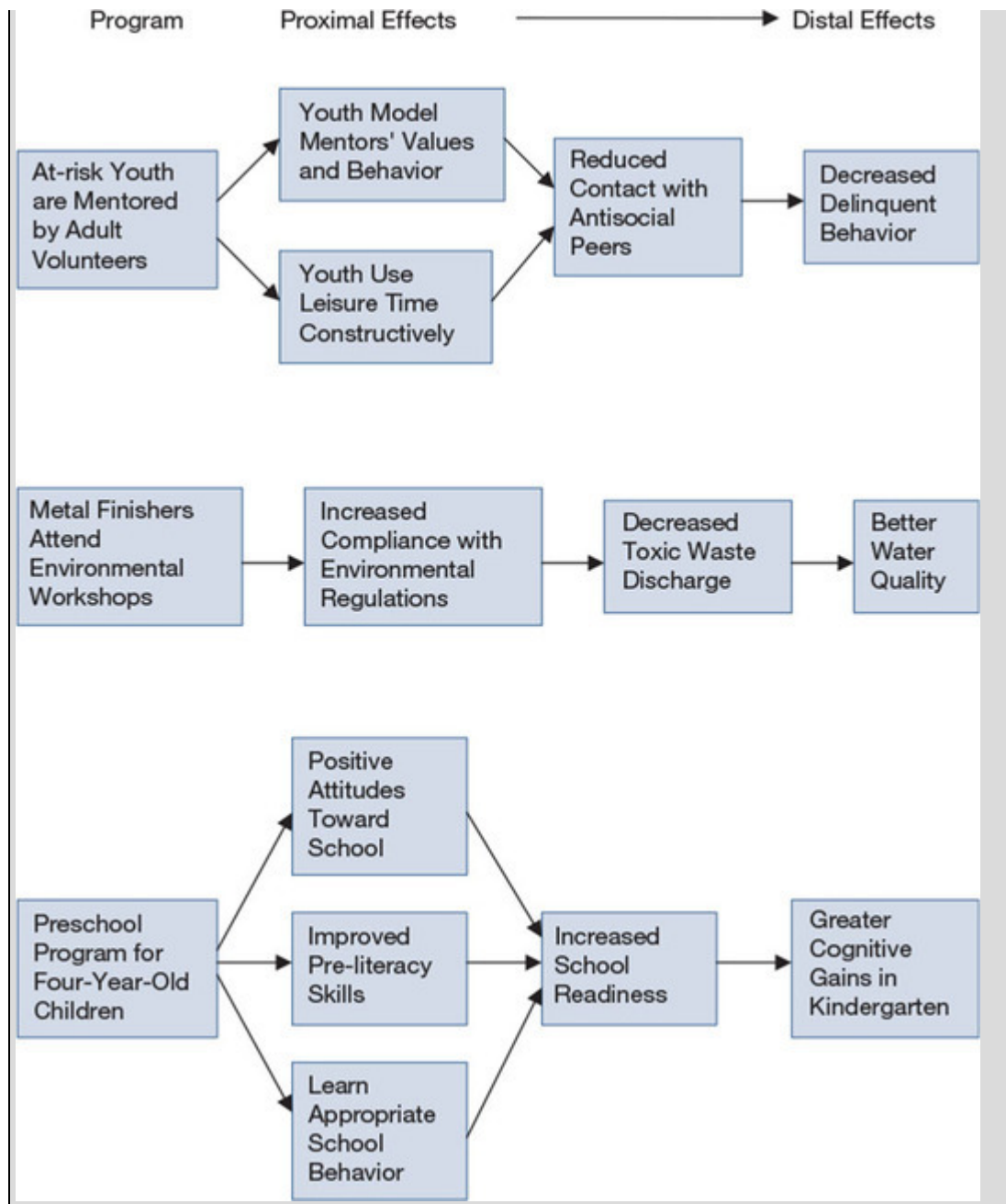
Exhibit 5-C shows several examples of the portion of program logic models that describes the impact theory (additional examples are in [Chapter 3](#)). For purposes of outcome assessment, it is useful to recognize the different character of the more proximal and more distal outcomes in these sequences. Proximal outcomes are those that program services are expected to affect most directly and immediately. These can be thought of as the "take away" outcomes: those program participants experience as a direct result of their participation and take with them out the door as they leave. For most social programs, these proximal outcomes are psychological: attitudes, knowledge, awareness, skills, motivation, behavioral intentions, and other such conditions that are susceptible to relatively direct influence by a program's services.

Proximal outcomes are rarely the ultimate outcomes the program intends to influence, as can be seen in the examples in [Exhibit 5-C](#). In this regard, they are not the most important outcomes from a social or policy perspective. However, this does not mean that they should be overlooked in any evaluation. These outcomes are the ones the program has the greatest capability to affect, so it can be informative to know whether they show evidence of program effects. If the program fails to influence these most immediate and direct outcomes, and the program theory is correct, then the more distal outcomes in the sequence are unlikely to occur. In addition, the proximal outcomes are generally the easiest to measure and the easiest to

assess for program effects. If the program is successful at influencing these outcomes, it is appropriate for it to receive credit for doing so. The more distal outcomes, which may be more difficult to measure, are also typically the ones most difficult to assess for program effects. Impact evaluation estimates of program effects on the distal outcomes will be more balanced and interpretable if information is also available on the proximal outcomes.

Nonetheless, it is the more distal outcomes that are usually the ones of greatest practical and policy importance. It is thus especially important to clearly identify and describe those distal outcomes that can reasonably be expected to be affected by the program. Generally, however, a program has less direct influence on the distal outcomes than on the proximal ones because the distal outcomes are typically influenced by many more factors extraneous to the program. This circumstance makes it especially important to define the distal outcomes in a way that aligns as closely as possible with the aspects of the social conditions program activities can plausibly affect. Consider, for instance, a tutoring program for elementary school children that focuses mainly on reading with the intent of increasing educational achievement. The educational achievement outcomes defined for an evaluation of this program should distinguish between those outcomes closely related to the reading skills the program teaches and other outcomes, such as mathematics, that are less likely to be influenced by what the program is actually doing.

Exhibit 5-C Examples of Program Impact Theories Showing Expected Program Effects on Proximal and Distal Outcomes



Prior Research

In identifying and defining outcomes, the evaluator should thoroughly examine prior research related to the program being evaluated, especially evaluation research on similar programs. Learning which outcomes have been examined in other studies may call attention to relevant outcomes that might otherwise be overlooked. It will also be informative to see how various outcomes have been defined and measured in prior research. In some cases, there may be relatively standard definitions and measures that, if adopted for the evaluation, would allow direct comparisons of the evaluation results with those reported for other programs. In other cases, there may be known problems with certain definitions or measures that the evaluator should be aware of.

Unintended Effects

So far, we have been considering how to identify and define the outcomes stakeholders expect the program to influence and those that are evident in the program's impact theory. There may be significant unintended effects of a program, however, on outcomes that are not identified through these means. Such effects may be positive or negative, but their distinctive character is that they emerge through some process that is not part of the program's design and direct intent. That feature, of course, makes them difficult to anticipate. Accordingly, the evaluator must often make special efforts to identify any outcomes outside the domain of those the program intends to affect that could be significant for a full understanding of the program's effects on the social conditions it addresses.

Prior research can often be especially useful on this matter. There may be outcomes other researchers have discovered in similar circumstances that can alert the evaluator to possible unanticipated program effects. In this regard, it is not only other evaluation research that is relevant but also any research on the dynamics of the social conditions in which the program intervenes. Research about the development of drug use and the lives of users, for instance, may provide clues about possible responses to a program intervention that the program plan has not taken into consideration.

Often, good information about outcomes on which there may be unintended effects can be found in the firsthand accounts of persons in a position to observe those effects. For this reason, as well as others mentioned elsewhere in this text, it is important for the evaluator to have substantial contact with program personnel at all levels, program participants, and other key informants with perspectives on the program and its effects. If unintended effects are at all consequential, there should be someone in the system who is aware of them and who, if asked, can alert the evaluator. These individuals may not present this information in the language of unintended effects on particular outcomes, but their descriptions of what they see and experience in relation to the program will be interpretable if the evaluator is alert to the possibility that there could be important program

effects not articulated in the program logic or intended by the key stakeholders.

Features of some programs can predictably raise concerns about unanticipated effects. New programs focused on specific outcomes can crowd out the time and resources formerly dedicated to influencing other outcomes. For example, this would occur if a new reading program in an elementary school reduced instruction time in mathematics and science with corresponding effects on achievement test scores for those subjects. Or the personnel required to operate a program might be drawn from other similar programs with consequences for the effects of those other programs on their intended outcomes. As part of a recent evaluation of a program to improve the lowest performing schools, for instance, Kho, Henry, Zimmer, and Pham (2018) found that the bonuses and additional pay offered to incentivize effective teachers to move to low-performing schools produced negative effects on student achievement in the schools that lost teachers. Another possibility is that place-based programs to reduce risky behaviors will simply displace those engaged in the behaviors to other locations. For example, cameras in parking decks to deter breaking into cars might move the car break-ins to locations out of range of the cameras. Reductions in the number of reported break-ins in the parking decks with cameras thus may be offset by increases in auto break-ins elsewhere.

Measuring Program Outcomes

Not every outcome identified through the procedures we have described will be of equal importance or relevance, so the evaluator does not necessarily need to measure all of them in order to conduct an evaluation. Some prioritization and selection may be appropriate. In addition, some relevant outcomes—for example, very long term ones—may be difficult or expensive to measure for practical reasons and, consequently, may not be feasible to include in an evaluation.

Once the relevant outcomes have been chosen and a full and careful description of each is in hand, the evaluator must face the issue of how to measure them. Outcome measurement is a matter of representing the circumstances defined as the outcome by means of observable indicators that vary systematically with changes or differences in those circumstances. Some program outcomes have to do with relatively simple and easily observed circumstances that are virtually one-dimensional. One outcome an industrial safety program may intend to affect, for instance, might be whether workers wear their safety goggles in the workplace. An evaluator can measure this outcome quite well for each worker at any given time with a simple observation and recording of whether the goggles are being worn and, by making periodic observations, extend the measurement to how frequently they are worn.

Many important program outcomes, however, are not as simple to measure as whether a worker is wearing safety goggles. To fully represent an outcome, it may be necessary to view it as multidimensional and differentiate multiple facets of it that are relevant to the effects the program is attempting to produce. [Exhibit 5-D](#), for instance, provides a description of juvenile delinquency as an outcome variable in terms of legally chargeable offenses committed. The chargeable delinquent offenses committed by juveniles, however, have several distinct dimensions that could be affected by a program attempting to reduce delinquency. To begin with, both the frequency of offenses and the seriousness of those offenses are likely to be relevant. Program personnel would not be happy to discover that they had reduced the frequency of offenses, but those still committed were now much more serious. Similarly, the type of offense may require

consideration. A program focusing on drug abuse, for example, may expect drug offenses to be the most relevant outcome, but it may also be sensible to examine property offenses because drug abusers may commit those to support their drug purchases. Other offense categories may be relevant, but less so, and it would obscure important distinctions to lump all offense types together as a single outcome measure and increase the possibility that lack of effects on the less relevant outcomes would mask the effect on property offenses.

Most outcomes are multidimensional in this way; that is, they have various facets or components the evaluator may need to take into account. The evaluator generally should think about outcomes as comprehensively as possible to ensure that no important dimensions are overlooked. This does not mean that all must receive equal attention or even that all must be included in the coverage of the outcome measures selected. The point is, rather, that the evaluator should consider the full range of potentially relevant dimensions before determining the final measures to be used.

[Exhibit 5-D](#) presents several other examples of outcomes, with various aspects and dimensions broken out.

One implication of the multiple dimensions of program outcomes is that a single outcome measure may not be sufficient to represent their full character. In the case of juveniles' offenses, for instance, the evaluation might use measures of offense frequency, severity, time to first offense after intervention, and type of offense as a battery of outcome measures that attempt to fully represent this outcome. Indeed, multiple measures of important program outcomes help the evaluator guard against missing an important program accomplishment because of a narrow measurement strategy that leaves out relevant outcome dimensions.

Exhibit 5-D Examples of Multiple Dimensions and Facets of Outcomes

- Juvenile delinquency
 - Number of chargeable offenses committed during a given period
 - Severity of offenses
 - Type of offense: violent, property crime, drug offenses, other
 - Time to first offense from an index date
 - Official response to offense: police contact or arrest; court adjudication, conviction, or disposition
- Toxic waste discharge

- Type of waste: chemical, biological; presence of specific toxins
- Toxicity, harmfulness of waste substances
- Amount of waste discharged during a given period
- Frequency of discharge
- Proximity of discharge to populated areas
- Rate of dispersion of toxins through aquifers, atmosphere, food chains, and the like
- School performance
 - Proficiency rates on standardized achievement tests by subject
 - School value-added scores
 - Chronic student absenteeism
 - Exclusionary discipline
 - Turnover of effective teachers

Diversifying measures can also safeguard against the possibility that poorly performing measures will underrepresent program effects and, by not measuring the aspects of the outcome a program most affects, make the program look less effective than it actually is. For outcomes that depend on observation, for instance, having more than one observer may be useful to avoid the biases associated with any one of them. An evaluator assessing children's aggressive behavior with their peers might want the parents' observations, the teachers' observations, and those of any other persons in a position to see a significant portion of the children's behavior. An example of multiple measures is presented in [Exhibit 5-E](#).

Multiple measures of important outcomes thus can provide broader coverage of the different facets of those outcomes and allow the strengths of one measure to compensate for the weaknesses of another. It may also be possible to statistically combine multiple measures into a single, more robust and valid composite measure that is better than any of the individual measures taken alone. In a program to reduce family fertility, for instance, changes in desired family size, adoption of contraceptive practices, and average desired number of children might all be measured and used in combination to assess the critical program outcome. Even when measures must be limited to a smaller number than comprehensive coverage might require, it is useful for the evaluator to elaborate all the dimensions and variations in order to make a thoughtful selection from the feasible alternatives.

The Norwegian Institute for Alcohol and Drug Research evaluated an initiative targeting the use of and harm from alcohol in six communities. The initiative was funded by the Directorate of Health and coordinated through regional centers. It emphasized selection and implementation of evidence-based strategies that targeted adolescents, their parents, and businesses selling alcohol and were oriented toward emphasizing delay of first alcohol use and reduced access and use. The evaluation used two different approaches to measuring alcohol use and access outcomes: (a) school-based surveys of 13- to 19-year-old students and (b) sending younger looking 18-year-olds to attempt to purchase beer in grocery stores.

Outcomes Measured

Students reported on drinking behavior during the past 12 months, including

- whether they had ever drunk alcohol,
- drinking frequency,
- whether they had ever been intoxicated, and
- intoxication frequency.

Students reported whether they had experienced alcohol-related harm during the past 12 months, including whether they had

- been in a fight,
- committed vandalism,
- driven a vehicle while under the influence of alcohol, and
- drunk so much that they had vomited.

Students reported on the availability of alcohol, including

- the frequency of procuring alcohol at off-premises outlets,
- the frequency of procuring alcohol at on-premises outlets, and
- the frequency of having been denied purchase at off-premises outlets.

Evaluators observed

- the frequency with which underage-appearing adolescents successfully purchased beer at grocery stores.

The evaluators reported no changes in the outcomes before and after the interventions were implemented. For example, approximately 50% of the youth who appeared to be underage were able to purchase beer in the grocery stores before and after the program. In addition to funding and implementation delays, the evaluators concluded, “Despite an initial emphasis on evidence-based strategies, a review of the relevant literature showed that few of the recommended strategies had any documented effects on drug use or related harm. A closer look at the literature regarding these strategies revealed that ‘evidence’ of effectiveness was limited.”

Source: Adapted from Rossow, Stordvoll, Baklien, and Pape (2011).

Measurement Procedures and Properties

Data on program outcomes have relatively few basic sources: observations, records, responses to interviews and questionnaires, standardized tests, physical measurement, and the like. The information from such sources becomes measurement when it is operationalized, that is, generated through a set of specified, systematic operations or procedures. The measurement of many outcome variables in evaluation uses procedures and instruments that are already established and accepted for those purposes in the respective program areas. This is especially true for the more distal and policy relevant outcomes. In health care, for instance, morbidity and mortality rates and the incidence of disease or particular health problems are measured in relatively standardized ways that differ mainly according to the nature of the health problem at issue. Academic performance is conventionally measured with standardized achievement tests and grade point average. Occupations and employment status ordinarily are assessed by means of measures developed by the Bureau of Labor Statistics.

For other outcomes, various ready-made measurement instruments or procedures may be available, but with little consensus about which are most appropriate for evaluation purposes. This is especially true for psychological outcomes such as depression, self-esteem, attitudes, cognitive abilities, and anxiety. In these situations, the task for the evaluator is generally to make an appropriate selection from the options available. Practical considerations, such as how the instrument is administered and how long it takes, must be weighed in this decision. The most important consideration, however, is how well a ready-made measure matches what the evaluator wants to measure. Having a careful description of the outcome to be measured, as illustrated in [Exhibit 5-B](#), will be helpful in making this determination. It will also be helpful if the evaluator has differentiated the distinct dimensions of the outcomes that are relevant, as illustrated in [Exhibit 5-D](#).

When ready-made measurement instruments are used, it is especially important to ensure that they are suitable for adequately representing the outcome of interest. A measure is not necessarily appropriate just because the name of the instrument, or the label given for the construct it measures,

is similar to the label given the outcome of interest. Different measurement instruments for the “same” construct (e.g., self-esteem, environmental attitudes) often have rather different content and theoretical orientations that give them a character that may or may not match the program outcome of interest once that outcome is carefully described. Convenience and familiarity are not sufficient criteria for selecting a measure. In a recent study of the validation of measures of the effects of teacher training programs, most measures, including observational ratings of student teaching by university supervisors, survey of teacher candidates’ dispositions, or ratings of portfolios of teacher candidates’ work were unrelated to their subsequent performance as teachers (Henry et al., 2013). Only college grade point average and number of math courses were systematically related to the teacher candidates’ effectiveness in the classroom.

For many of the outcomes of interest to evaluators, there are neither established measures nor a range of ready-made measures from which to choose. In these cases, the evaluator must develop the measures. Unfortunately, there is rarely sufficient time and resources to do this properly. Some ad hoc measurement procedures, such as extracting specific relevant information from administrative records, are sufficiently straightforward to qualify as acceptable measurement practice without further demonstration. All other measurement procedures, however, such as questionnaires, attitude scales, knowledge tests, and observational coding schemes, are not as straightforward as administrative data. Constructing such measures so that they adequately measure the critical outcomes in the program impact theory in a consistent fashion is often not easy. Because of this, there are well-established measurement development procedures for doing so (see, e.g., Bastian, Henry, Pan, & Lys, 2016) that involve technical considerations and generally require a significant amount of testing, analysis, revision, and validation before a newly developed measure can be used with confidence. When an evaluator must develop a measure without going through these steps and checks, the resulting measure may be reasonable on the surface but will not necessarily perform well for purposes of validly and reliably measuring program outcomes.

When ad hoc measures must be developed without the opportunity to do so in a systematic and technically proper manner, it is especially important that their basic measurement properties be checked before weight is put on them in an evaluation. Indeed, even in the case of ready-made measures and accepted procedures for assessing certain outcomes, it is wise to confirm that the respective measures perform well for the situation in which they will be applied. There are three measurement properties of particular concern in this regard: reliability, validity, and sensitivity.

Reliability

The **reliability** of a measure is the extent to which it produces the same results when used repeatedly to measure something that has not changed. Variation in the results constitutes measurement error. So, for example, a postal scale is reliable to the extent that it reports the same weight for the same envelope when it is weighed more than once. No measuring instrument, classification scheme, or counting procedure is perfectly reliable, but different types of measures have varying degrees of reliability problems. Measurements of physical characteristics for which standard measurement devices are available, such as height and weight, will generally be more consistent than measurements of psychological characteristics, such as intelligence measured with an IQ test.

Performance measures, such as standardized achievement tests, in turn, have been found to be more reliable than measures relying on recall, such as reports of household expenditures for consumer goods. For evaluators, a major source of unreliability lies in the nature of measurement instruments that are based on participants' responses to written or oral questions posed by researchers. In such situations, reliability implies that two individuals whose outcomes are the same would be assigned the same value on the outcome measure, and individuals whose outcomes are different would be assigned different values on the outcome measure. Differences in the testing or measuring situation, observer or interviewer differences in the administration of the measure, and variation in respondents' recall or engagement in the measurement process will contribute to unreliability.

The effect of measurement unreliability is to dilute and obscure real differences. A truly effective intervention, the outcome of which is measured unreliably, will appear to be less effective than it actually is. More reliable measures make estimates of average outcomes more precise and therefore make it easier to distinguish real change in these averages from chance variation. However, there are no hard and fast rules about acceptable levels of reliability. The extent to which measurement error can obscure a meaningful program effect on an outcome depends in large part on the

magnitude of that effect and the size of the sample with which the effect is estimated (matters that are discussed more fully in [Chapter 9](#)).

The most straightforward way for the evaluator to check the reliability of a candidate outcome measure is to administer it at least twice under circumstances when the outcome should not change in between.

Technically, the conventional index of this *test-retest* reliability is a statistic known as the product-moment correlation between the two sets of scores, which varies between 0.00 and 1.00 for a test-retest application. For many outcomes, however, this check is difficult to make because the outcome may change naturally between measurement applications that are not closely spaced. For example, questionnaire items asking students how well they like school may be answered differently a month later, not because the measure is unreliable but because intervening events have made the students feel differently about school. When the measure involves responses from people, on the other hand, administering it at closely spaced intervals will yield biased results to the extent that respondents remember and repeat their prior responses rather than generating fresh ones. When the measurement cannot be repeated before the outcome changes, reliability is usually checked by examining the consistency among similar items in a multi-item measure administered at the same time (referred to as internal consistency reliability and indexed with a statistic called Cronbach's alpha).

For many of the ready-made measures evaluators use, reliability information will be available from other research or from reports of the original development of the measure. Reliability can vary according to the sample of respondents and the circumstances of measurement, however, so it is not always safe to assume that a measure that has been shown to be reliable in other applications will be reliable when used in a particular evaluation.

Validity

The issue of measurement validity is more difficult than the problem of reliability. The **validity** of a measure is the extent to which it measures what it is intended to measure. For example, juvenile arrest rates provide a valid measure of delinquency only to the extent that they accurately reflect how frequently the juveniles have engaged in chargeable offenses. To the extent that they also reflect police arrest practices, their validity as measures of the delinquent behavior of the juveniles is compromised.

Although the concept of validity and its importance are easy to comprehend, it is usually difficult to test whether a particular measure is valid for the characteristic of interest. With outcome measures used for evaluation, validity turns out to depend very much on whether a measure is accepted as valid by the appropriate stakeholders (often referred to as *face validity*). Confirming that it represents the outcome intended by the program when that outcome is fully and carefully described (as discussed earlier) can provide some assurance of validity for the purposes of the evaluation. Using multiple measures of the outcome in combination can also provide some protection against the possibility that any one of those measures does not tap into the actual outcome of interest.

Empirical demonstrations of the validity of a measure depend on some comparison that shows that the measure yields the results expected if it were, indeed, valid. For instance, when the measure is applied along with alternative measures of the same outcome, such as those used by other evaluators, the results should agree to a reasonable order of approximation. Similarly, when the measure is applied to situations recognized to differ on the outcome at issue, the results should differ. Thus, a measure of environmental attitudes should sharply differentiate members of the local Sierra Club from members of an off-road dirt bike association. Validity is also demonstrated by showing that results on the measure relate to or predict other characteristics expected to be related to the outcome. For example, an examination of concurrent predictive validity could assess the extent to which an assessment of the planning skills exhibited in the portfolios of work submitted by teacher candidates correlates with their

supervisor's ratings of their planning skills. Another type of predictive validity is especially salient when measuring a program's short-term outcomes. Predictive validity of the short-term outcome measures occurs when these measures predict or are highly correlated with longer term outcomes.

Sensitivity

A primary function of outcome measures is to detect changes or differences in outcomes that represent program effects. To accomplish this well, outcome measures must be sensitive to such effects. The **sensitivity** of a measure is the extent to which the values on the measure change when there is a change or difference in the thing being measured. Suppose, for instance, that we are measuring body weight as an outcome for a weight-loss program. A finely calibrated scale of the sort used in physicians' offices might measure weight to within a few ounces and, correspondingly, be able to detect weight loss in that range. In contrast, the weigh-in-motion scales for trucks on interstate highways are also valid and reliable measures of weight, but they are not sensitive to differences smaller than a few hundred pounds. A scale that was not sensitive to meaningful fluctuations in the weight of the dieters in the weight-loss program would be a poor choice to measure that outcome.

There are two main ways in which the kinds of outcome measures frequently used in program evaluation can be insensitive to changes or differences of the magnitude the program might produce. First, the measure may include elements that relate to something other than what the program could reasonably be expected to change. These dilute the concentration of elements that are responsive and mute the overall response of the measure. Consider, for example, a math tutoring program for elementary school children that has focused on fractions and long division problems for most of the school year. The evaluator might choose the state's required math achievement test as a reasonable outcome measure. Such a test, however, will include items that cover a wider range of math problems than fractions and long division. The children's higher scores on items involving fractions or long division might be obscured by their performance on other topics that were not addressed by the tutoring program but are averaged into the final score. A more sensitive measure would be one that included only the math content aligned with what the program actually covered.

Second, outcome measures may be insensitive to the kinds of changes or differences induced by programs when they have been developed largely

for diagnostic purposes, that is, to detect individual differences. The objective of measures of this sort is to spread the scores in a way that differentiates individuals who have more or less of the characteristic being measured. Most standardized psychological measures are of this sort, including, for example, personality measures, measures of clinical symptoms (depression, anxiety, etc.), measures of cognitive abilities, and attitude scales. These measures are generally good for determining who is high or low on the characteristic measured, which is their purpose, and thus are helpful for, say, assessing needs or problem severity. However, when applied to a group of individuals who differ widely on the measured characteristic before participating in a program, they may yield such a large variation in scores after participation that any increment of improvement produced by the program will be lost amid the differences between individuals. From a measurement standpoint, the individual differences to which these measures respond so well constitute irrelevant noise for purposes of detecting change or group differences and tend to obscure those effects. [Chapter 9](#) discusses some ways the evaluator can compensate for this source of measurement insensitivity.

The best way to determine whether a candidate outcome measure is sufficiently sensitive for use in an evaluation is to find research in which it was used successfully to detect a change or difference on the order of magnitude the evaluator expects from the program being evaluated. The clearest form of this evidence, of course, comes from evaluations of very similar programs in which significant change or differences were found using the outcome measure. Appraising this evidence must also take the sample size of the prior evaluation studies into consideration, because the size of the sample also affects the ability to detect differences (discussed in more detail in [Chapter 9](#)).

An analogous approach to investigating the sensitivity of an outcome measure is to apply it to groups of known difference, or situations of known change, and determine how responsive it is. Consider the example of the math tutoring program mentioned earlier. The evaluator may want to know whether the standardized math achievement tests administered by the state every year will be sufficiently sensitive to use as an outcome measure. This may be a matter of some doubt, given that the tutoring focuses on only a

few math topics, while the achievement test covers a wide range. To check sensitivity before using this test to evaluate the program, the evaluator might first administer the test to a classroom of children before and after they study fractions and long division. If the test proves sufficiently sensitive to detect changes over the period when only these topics are taught, it provides some assurance that it will be responsive to the effects of the math tutoring program when used in the evaluation. Also, in some situations like this it may be possible to identify the test items covering the program content, in this case fractions and long division, and extract them from the overall measure as the basis for a new measure that is better aligned to the program and more sensitive to its effects.

Choice of Outcome Measures

As the discussion so far has implied, selecting the best measures for assessing outcomes is a critical measurement problem in evaluation (Rossi, 1997). We recommend that evaluators invest the necessary time and resources to develop and test appropriate outcome measures ([Exhibit 5-F](#) describes an exemplary effort to develop and assess outcome measures in an intervention area in which negative outcomes had previously dominated). A poorly conceptualized outcome measure may not properly represent the goals and objectives of the program being evaluated, leading to questions about the validity of the measure. An unreliable or insufficiently sensitive outcome measure is likely to underestimate the effectiveness of a program and lead to incorrect conclusions about its impact. In short, a measure that is poorly chosen or poorly conceived can completely undermine the worth of an impact assessment by producing misleading estimates. The evaluator can have confidence that an outcome measure is capable of measuring actual program effects only if it is valid, reliable, and appropriately sensitive to change.

Exhibit 5-F Valid and Reliable Measures of Positive Development of Adolescents

On the basis of a conviction that positive measures of adolescent well-being were largely absent from evaluations of interventions to improve young people's development, Child Trends undertook the Flourishing Children Project. The purpose of the project was to develop and assess short, valid, and reliable measures of positive child well-being that would work with diverse adolescents and their parents and could be used cost-effectively in evaluations or surveys of this population.

The project team developed a large set of candidate items, then conducted interviews with adolescents to explore their relevance and salience for that population. After the most promising items for several distinct measurement scales were identified, they were pilot-tested in a nationally representative Web-based survey with adolescents between 12 and 17 years old and their parents. The resulting data were used to examine the concurrent validity, reliability, and distributional properties of the respective measurement scales. Two of those scales are described below.

Diligence and Reliability

Definition: “Performing tasks with thoroughness and effort from start to finish where one can be counted on to follow through on commitments and responsibilities. It includes working hard or with effort, having perseverance and performing tasks with effort from start to finish, and being able to be counted on.” Items included “Do you work harder than others your age?” and “Do you finish the tasks that you start?” The internal consistency reliability index (Cronbach’s alpha) was above .75, which is considered good. In terms of concurrent validity, diligent and reliable adolescents were less likely to smoke, get into fights, or report being depressed, and more likely to get good grades.

Initiative Taking

Definition: “The practice of initiating an activity toward a specific goal by adopting the following characteristics: reasonable risk taking and openness to new experiences, drive for achievement, innovativeness, and willingness to lead.” Items included “I like coming up with new ways to solve problems” and “I am a leader, not a follower.” The internal consistency reliability was above .70, which is considered acceptable. In terms of concurrent validity, initiative taking adolescents were less likely to smoke or report being depressed and more likely to get good grades.

Source: Adapted from Lippman et al. (2014).

Monitoring Program Outcomes

With adequate measures of significant program outcomes in hand, they can be used in various ways by evaluators or program managers to learn something about the performance of the program. The simplest application is outcome monitoring, defined as the regular measurement and reporting of indicators of the status of the individuals or social conditions with which the program has intervened. Outcome monitoring is similar to process monitoring, as described in [Chapter 4](#), with the difference that the information regularly collected and reviewed describes program outcomes rather than program process. Outcome monitoring for a job training program, for instance, might involve routinely telephoning participants 6 months after completion of the program to ask whether they are employed and, if so, what jobs they have and what wages they are paid. Detailed discussions of outcome monitoring (sometimes part of what is referred to as performance monitoring) and its relationship to program evaluation can be found in McDavid, Huse, and Hawthorn (2013), Kettner, Moroney, and Martin (2017), and Hatry (2014).

Outcome monitoring requires that measures be identified for important program outcomes that are practical to collect routinely and informative with regard to the performance of the program. The latter requirement is particularly difficult. As discussed earlier in this chapter, simple measurement of outcomes provides information only about the status or level of the outcome, such as the number of children in poverty, the prevalence of drug abuse, the unemployment rate, or the reading skills of elementary school students. That information by itself is not sufficient to identify change in the outcome or to link any change specifically to program effects.

The source of this limitation, as mentioned earlier, is that there are usually many influences on the outcomes of interest other than the efforts of the program. Thus, poverty rates, drug use, unemployment, reading scores, and so forth change for any number of reasons related to the economy, social trends, and the influence of other programs and policies. Isolating program effects in a convincing manner from such other influences requires the special techniques of impact evaluation discussed in [Chapters 6](#), [7](#), and [8](#).

All that said, outcome monitoring can provide useful, relatively inexpensive, and informative feedback that can help program managers better administer and improve their programs. The remainder of this chapter discusses the procedures, potential, and pitfalls of outcome monitoring.

Indicators for Outcome Monitoring

The outcome measures that serve as indicators for use in outcome monitoring schemes should be as responsive as possible to program influences. For instance, the outcome indicators should be measured on the members of the target population who actually received the program services. This means that readily available social indicators for the geographic areas served by the program, such as census data or regional health data, are less valuable choices for outcome monitoring if they include an appreciable number of persons not actually served by the program.

The most interpretable outcome indicators are those that involve characteristics or behaviors that only the program is likely to have affected to any appreciable degree. Consider, for instance, a city street-cleaning program aimed at picking up litter, leaves, and the like from the municipal streets. Photographs of the streets that independent observers rate for cleanliness would be informative for assessing the effectiveness of this program. Short of a small hurricane blowing all the litter into the next county, there simply is not much else likely to happen that will clean the streets. Also, proximal outcomes from the impact theory may be especially informative in this regard. For a smoking cessation program, for instance, familiarity with relapse prevention skills, when those skills were a focus of the program, is less likely to be influenced by factors extraneous to the program than the actual amount of smoking.

An informative indicator that can be easily linked to program experience is client satisfaction, increasingly called customer satisfaction even in human service programs. Although not technically a program outcome as defined early in this chapter, direct ratings by recipients of the benefits they believe the program provided them, or not, are useful feedback for a program. In addition, creating feelings of satisfaction about the interaction with the program among the participants is itself usually an important program accomplishment. The more pertinent information comes from participants' reports of whether, with the benefit of hindsight after program participation, they believe they received the specific benefits the program intended as a

result of the service delivered (see [Exhibit 5-G](#) for an example). The limitation of such indicators is that program participants may not always be in a position to recognize or acknowledge program benefits, or may be reluctant to appear critical and thus overrate them.

Pitfalls in Outcome Monitoring

Because of the dynamic nature of the social conditions that typical programs attempt to affect, the limitations of outcome indicators, and the pressures on program agencies, there are various pitfalls associated with program outcome monitoring. Thus, while outcome indicators can be a valuable source of information for program decision makers, they must be developed and used carefully.

One important consideration is that any outcome indicator to which program funders or other influential decision makers give serious attention will also inevitably receive emphasis from program staff and managers. If the outcome indicators are not appropriate or fail to cover important outcomes, efforts to improve the performance they reflect may distort program activities. In the movement for greater accountability for community colleges, for instance, graduation rates are a common performance indicator. However, a focus on those rates alone could provide an incentive for college administrators to put additional admission requirements in place to screen out applicants less likely to graduate, even though that would run against the open-admissions policies that are a hallmark of community colleges. In economics, these are called perverse incentives. They can be offset by including other high-priority performance indicators that counter those perverse incentives (e.g., an indicator prioritizing higher percentages of disadvantaged applicants among those who enroll in the community colleges).

A related problem is the corruptibility of indicators. This refers to the natural tendency for those whose performance is being evaluated to attempt to skew the indicators in a favorable direction. In a program for which the rate of postprogram employment among participants is a major outcome indicator, for instance, consider the pressure on program staff assigned to telephone participants and ask about their job status. Even with a reasonable effort at honesty, ambiguous cases will more likely than not be recorded as employment. It is usually best for such information to be collected by interviewers independent from the program. If it is collected internal to the

program, it is especially important that careful procedures be used and that the results be verified in some convincing manner.

Exhibit 5-G Clinic Patient Satisfaction With HIV Services

With the change in the natural course of HIV/AIDS resulting from the use of highly active antiretroviral therapy, individuals with HIV/AIDS are living longer and receiving ambulatory care for longer periods as well. Recognizing the importance of client satisfaction to the delivery of high-quality services, the largest ambulatory clinic in Australia set out to develop a multidimensional measure of client satisfaction and administer a survey using those measures. The measures and the survey responses are shown in the table below.

The clients were generally satisfied with the services and the personnel delivering services, except for wait time on arrival. However, client satisfaction varied for different subgroups. For example, clients involved with the clinic for shorter periods and those who visited the clinic less frequently were more satisfied. From qualitative interviews that were conducted alongside the surveys, the evaluators found that “good rapport [between the client and the health care provider] was the main reason for staying with the same [health care provider].”

Ranking	Aspect	Percentage of Satisfied Respondents	Mean Satisfaction Score (Range = 1–5)
1	Maintenance of confidentiality	97	4.45
2	Knowledge of HCP	96	4.74
3	Overall satisfaction	95	4.46
4	Attitude of HCP	94	4.73
5	HCP understood clients' needs	94	4.68
6	Physical environment (waiting areas)	93	4.22
7	Suitable appointment time	91	4.29
8	Management of staff (waiting areas)	91	4.21
9	Location	91	4.27
10	Advice given during consultation	90	4.53
11	Included clients' decisions in treatment	89	4.56
12	Length of consultation time	84	4.45
13	Uninterrupted consultation	83	4.31
14	Availability of HCP	83	4.28
15	Benefitted more than expected	79	4.15
16	Waiting time on arrival	34	2.74
Overall mean (excluding "Waiting time on arrival")		90	4.42
Overall mean (including "Waiting time on arrival")		86	4.32

Source: Adapted from Chow, Li, and Quine (2012).

Note: HCP = health care provider.

Source: Adapted from Chow, Li, and Quine (2012).

Note: HCP = health care provider.

Another potential pitfall has to do with the interpretation of the outcome indicator data. Given a range of factors other than program performance that may influence those indicators, interpretations made out of context can be misleading and, even with proper context, can be difficult. To provide suitable context for interpretation, outcome indicators must generally be accompanied by other information that provides a relevant basis for comparison or explanation. We discuss the kinds of information that can be helpful in the following section.

Interpreting Outcome Data

Outcome data collected through routine outcome monitoring can be especially difficult to interpret if not accompanied by information about changes in client mix, relevant demographics, local economic trends, and the like. Job placement rates, for instance, are more accurately interpreted as a program performance indicator in light of information about the seriousness of participants' unemployment problems and the local job market. A low placement rate may not reflect poorly on program performance if the program is working with clients who have few job skills and long unemployment histories in an economy with scarce job openings for low-skilled workers.

Similarly, outcome data usually are more interpretable when accompanied by information about program process and service utilization. The job placement rate for clients completing training may look favorable but, nonetheless, be a matter for concern if, at the same time, the rate of training completion is low. The favorable placement rate may have resulted because all the clients with serious problems dropped out, leaving only the cream of the crop for the program to place. It is especially important to incorporate process and utilization information in the interpretation of outcome indicators when comparing different units, sites, or programs. It would be neither accurate nor fair to form a negative judgment of one program unit that was lower on an outcome indicator than other units without considering whether it was dealing with more difficult cases, maintaining lower dropout rates, or coping with other extenuating factors.

Indeed, the greatest utility of an outcome monitoring scheme for program managers is likely to come from cross-indexing outcome data with data on selected indicators of program performance in the context of background data on extraneous factors also likely to influence the outcomes. This can be especially informative with a focus on variation over time. The outcomes being monitored will almost always vary across the times at which they are measured. When a key outcome indicator rises or falls, a manager might first look for corresponding change in the profile of program performance indicators.

If, when the outcome improves, program performance has also improved, and when the outcome drops, program performance has declined as well, the manager has some basis for believing that the program makes a difference. Even more important, this pattern provides guidance for program improvement. It would then be informative and reassuring to a manager if, when improvements were made, outcomes improved as well. If there is no correspondence between variation in program performance and variation in the outcome measures, it may be because the program is functioning continuously at a high (or a low) level. But it may also be that the various indicators have not been chosen well, or even that the program, in fact, has little influence on the outcome.

Of course, variation across time in outcome indicators can be the result of extraneous factors outside the program's control. That is why it is relevant for an outcome monitoring scheme to also include data on a selected set of such factors. Perhaps most important is intake data for the program participants whose outcomes are being monitored. Variation across time in the severity of the issues they bring to the program, relevant demographics, amenability to intervention, and other such characteristics can be expected to affect outcomes. That variation, therefore, needs to be monitored and taken into account when interpreting any relationship between change on program performance indicators and change on outcomes measures. Similar considerations apply when there are relevant changes or trends in the local environment likely to influence the outcomes. With awareness of the multiple factors at play, a comprehensive program process and outcome monitoring dashboard can be a very useful tool for managers striving to maintain a high level of program performance or to improve performance to attain better outcomes. [Exhibit 5-H](#) describes a data dashboard that has many of these favorable characteristics, though not all of them are depicted there.

It may also be informative for interpretation of outcome monitoring data if they are broken out separately for subgroups of clients of particular concern to the program and/or who have characteristics at intake expected to relate to their success. Without that disaggregation, especially good or poor outcomes for such groups might be masked in the overall outcome results. Also, for program improvement purposes, managers may want to direct

particular attention to groups with poorer outcomes, or add supplementary services for them, and track whether any response to those efforts shows up in later outcome data.

Another potentially informative configuration of data from outcome monitoring, when applicable, is to compare outcome status with preprogram status on the same outcome measure (e.g., from intake data), for the same program participants. This will reveal the amount of change that has taken place for each cohort of participants, and variation in that change can be tracked across successive cohorts. For example, it is less informative to know that 40% of the participants in a job training program are employed 6 months afterward than to know that this represents a change from a preprogram status in which 90% had not held a job during the previous year. One interpretive aid for such pre-post comparisons is to define a reasonable success threshold and track the proportions that move from below that threshold to above it after receiving service. Thus, if the threshold is defined as “holding a full-time job continuously for 6 months,” the proportion of participants falling below that threshold for the year prior to program intake and the proportion above that threshold during the year after completion of services could be examined.

The main drawback to simple pre-post (before and after) comparisons is that any improvement they reveal cannot be confidently ascribed to program effects. One of the main reasons people choose to enter job training programs, for instance, is that they are unemployed and experiencing difficulties obtaining employment. Hence, they are at a low point at the time of entry into the program, and their situation from there is more likely to improve than deteriorate with or without the assistance of the program. Pre-post comparisons for such programs thus almost always show an upward trend that may have little to do with program effects. Other factors can also influence pre-post change, for instance, an improvement in the job market. In general, then, while pre-post comparisons may provide useful feedback to program administrators, they do not usually provide credible findings about a program’s impact. The rare exception is when there are no intervening events or trends that might plausibly account for a pre-post difference. This is unlikely when the human organism is involved, but may characterize some physical systems. Measures of radon or lead in

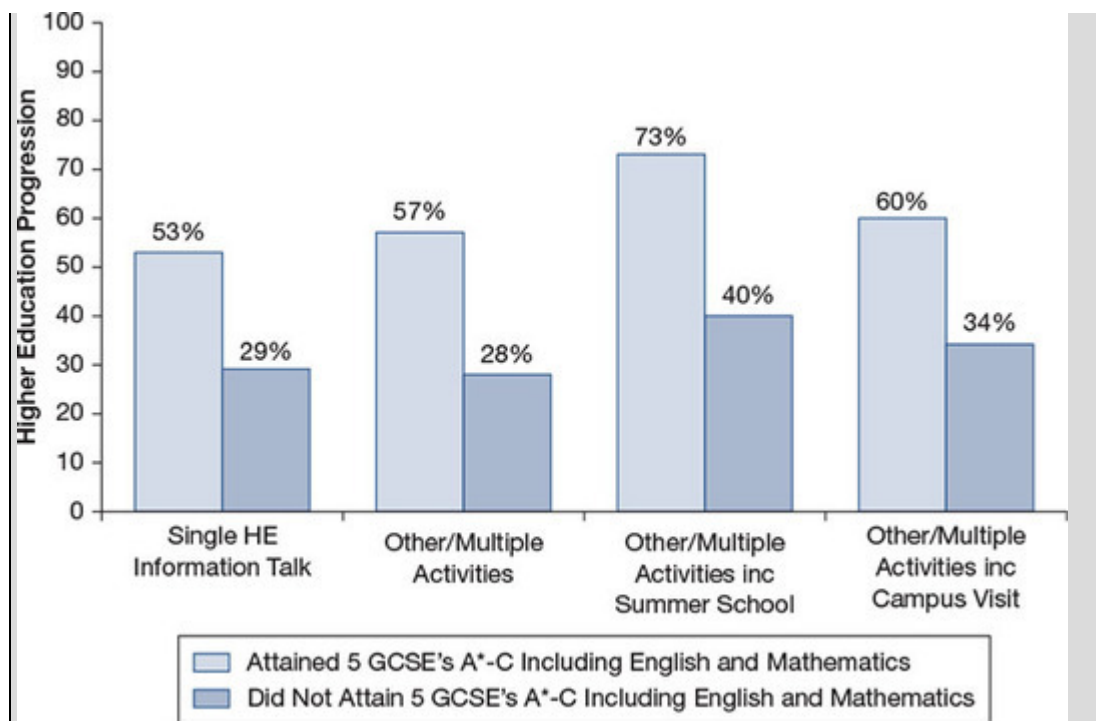
the paint in low-income housing in the context of abatement programs, for example, are situations in which pre-post comparisons may largely reflect program effects given few other influences likely to affect them.

Exhibit 5-H Monitoring Higher Education Outreach Interventions in England Using the Higher Education Access Tracker

In England, several groups of young people are underrepresented in the nation’s colleges and universities, including White working-class men, Black and ethnic minority students, and students from low-income backgrounds. Colleges and universities have been tasked by the government to reach out and engage with these underrepresented students and support their progression to higher education. These activities include, for example, providing information about higher education finance and progression routes, hosting summer schools on university campuses, and offering campus visits. In tandem with these activities, more than 70 higher education institutions have joined a collaborative initiative known as HEAT (Higher Education Access Tracker) that provides information institutions can use to monitor their activities and outcomes.

The graphic below provides an example of the HEAT data dashboard showing the number of activities institutions have added to the database with recorded contact hours, registered students, and the number of student records with incomplete data, including the types of data that are missing. HEAT also provides institutions with ongoing outcome data and infographics. For example, the graphic below shows the percentage of students who progressed to an English institution of higher education contextualized by the types and amount of interventions they participated in and their prior educational history. This graphic demonstrates that students who participated in the most intensive interventions—those that included multiple activities and a summer school—were those most likely to progress to higher education, even among the students with weaker educational backgrounds (A-C examination results at age 16).

Figure 1 Higher Education Progression Rate by GCSE Attainment and Outreach Engagement



Source: Higher Education Access Tracker (2017).

Note: GCSE = General Certificate of Secondary Education.

The information generated by outcome monitoring schemes will be available not only to program managers, but generally to program sponsors and other stakeholders as well. They will inevitably interpret the outcome data in relation to their expectations for the impact of the program on those outcomes. It is important that they understand that such outcome data are not direct reflections of program effects and, indeed, may be very misleading about actual program effects. Extreme outcomes may not cause much confusion. For instance, suppose that several months after a program to treat alcoholism, more than 90% of the participants were no longer drinking. Given the typically high relapse rates for this population, that's a remarkable outcome that the program quite likely influenced. On the other hand, if only 10% have stopped drinking, there's good reason to question the effectiveness of the program.

In reality, of course, the observed outcomes would probably be more ambiguous, say 45% still drinking. This more likely middle ground requires caution about any interpretation that attempts to make attributions about program effects. Further information, which may not be available, is

required before any conclusion can be drawn about the program's influence on that outcome. Consistent patterns of covariation over time between program performance indicators and outcomes, as described above, might support somewhat stronger conclusions. Another approach would be to compare the program's outcomes with those from similar programs, a tactic known as *benchmarking* (Keehley, Medlin, Longmire, & MacBride, 1997). This is most informative when the comparison is with a very similar program that serves a very similar clientele, especially if there is reason to believe that the comparison program is one that is especially effective.

The broader theme inherent in this discussion of outcome monitoring, however, is that it should not be viewed or used as a scheme for assessing the effects of a program on the respective outcomes. Its main value is as a management tool that informs program decision makers about how well program participants are doing after they leave the program, which subgroups are doing better or worse, and how these aspects of the outcome picture are changing over time. Most important, thoughtful use of the data from a well-developed outcome monitoring scheme can provide valuable guidance to efforts to improve the program and feedback about the results.

Summary

- Programs are designed to affect some problem or need in positive ways. The characteristics or behaviors of the target population or social conditions that are the targets of those efforts to bring about change constitute the relevant outcomes for the program.
- Identifying outcomes relevant to a program requires input from stakeholders, review of program documents, and articulation of the impact theory embodied in the program's logic. Evaluators should also attend to relevant prior research and consider possible unintended outcomes.
- Outcome measures can describe the status of the individuals or other units that constitute the target population whether or not they have participated in the program. They can also be used to describe change in outcomes over time and are used in impact evaluation designs that attempt to determine a program's effect on relevant outcomes.
- Because outcomes are affected by events and experiences that are independent of a program, changes in the levels of outcomes cannot be directly interpreted as program effects.
- To produce credible results in any evaluation application, outcome measures need to be reliable, valid, and sensitive to the order of magnitude of change that the program might be expected to produce. In addition, it is often advisable to use multiple measures or outcome variables to reflect multidimensional outcomes and to correct for possible weaknesses in one or more of the measures.

- Outcome monitoring schemes track selected outcomes over time and can serve program managers and other stakeholders by providing timely and relatively inexpensive descriptive information. Carefully used, that descriptive information can be useful for guiding efforts to improve programs.
- The interpretation of data from outcome monitoring requires consideration of a program's environment, events taking place during the program, the characteristics of the participants, and various other factors with the potential to influence the selected outcome measures. Those data will say little about a program's effects on the outcomes, but can help differentiate the influence of the program on the outcomes of interest from extraneous influences on those outcomes.

Key Concepts

Impact 119
Outcome 116
Outcome change 117
Outcome level 117
Program effect 117
Reliability 128
Sensitivity 130
Validity 129

Critical Thinking/Discussion Questions

1. Define an outcome. What makes an outcome different from an output? Explain outcome level, outcome change, and program effect. What are the differences in the kinds of information provided to program stakeholders by measures of these different aspects of outcomes?
2. Explain four ways relevant outcomes for a given program can be identified.
3. What are five areas of concern in measuring program outcomes? How are they related, and how can an evaluator attempt to deal with each area of concern in conducting an evaluation?

Application Exercises

1. Locate a Web site for a social program. Review the services that program delivers and the stated goals and objectives of the program. Taking that information at face value, identify three specific outcomes you would measure as a part of an evaluation of this program. Describe how you would measure each of these outcomes.
2. Benchmarking is described in this chapter as the process by which an evaluator compares the program's outcomes with those from similar programs. Using the social program in Exercise 1, locate a study that could be used for benchmarking. Summarize the study's findings and describe the benchmarks you would use in your evaluation.

Chapter 6 Impact Evaluation Isolating the Effects of Social Programs in the Real World

[The Nature and Importance of Impact Evaluation](#)

[Additional Impact Questions](#)

[When Is an Impact Evaluation Appropriate?](#)

[What Would Have Happened Without the Program?](#)

[The Logic of Impact Evaluation: The Potential Outcomes Framework](#)

[The Fundamental Problem of Causal Inference: Unavoidable Missing Data](#)

[The Validity of Program Effect Estimates](#)

[Summary](#)

[Key Concepts](#)

In the eyes of many evaluators and policymakers, impact evaluations answer one of the most important questions about a social program: Did the program make the intended beneficiaries better off? However, the reality of social programs and the nature of their effects challenge the ability of impact evaluators to answer this question definitively. In this chapter, we lay out the logic and the challenges of impact evaluation. Central to the logic as well as the challenges is determining what would have happened in the absence of the program to contrast with the actual outcomes for program participants. Understanding the importance of answering that question convincingly and what is required to do so is critical to conducting a valid impact evaluation.

With rare cynical exceptions, policymakers and sponsors launch programs with the intent of bringing about beneficial changes in some condition deemed undesirable. That is, the program is expected to produce better outcomes than would occur without the program. The difference between the outcomes that occur with implementation of the program and those that would have occurred otherwise is the **program effect** or, as it is often called, the **program impact**. Every program interjected into the social fabric perturbs it in some way, whether in the intended way or not and whether trivial or consequential. We can thus distinguish between the outcomes the program targets for improvement and any other outcomes,

beneficial or otherwise, that the program may also influence. What is often called the law of unintended consequences alerts us to be especially mindful of the latter.

The Nature and Importance of Impact Evaluation

Because it addresses the primary purpose of a program, questions about program impact are typically central to the concerns of program sponsors, advocates, critics, and potential beneficiaries. Thus, among the types of evaluations presented in [Chapter 1](#), **impact evaluation** is one of the most highly valued by stakeholders and evaluators alike, in no small measure because of its potential to influence policy and high-level program decisions. Indeed, it would be difficult to overstate the importance of impact evaluation and its prominence among the various types of evaluation. In some disciplines, such as economics, program evaluation is synonymous with impact evaluation, and training in program evaluation focuses exclusively on methods for determining impact and their application to various program circumstances.

Identifying and measuring program effects is a matter of demonstrating that the program has caused some change in the outcomes of the individuals exposed to the program that would not otherwise have occurred. Fundamentally, then, impact evaluation deals with cause-and-effect relationships. In the social sciences, causal relationships are ordinarily understood in terms of probabilities. Thus, the statement “A causes B” means that if we introduce A, B is more likely to result than if we do not introduce A, all else equal. This statement does not imply that B always results from A, nor does it mean that B occurs only if A happens first. To illustrate, consider a job training program designed to reduce unemployment. If successful, it will increase the probability that unemployed participants will subsequently be employed. Even a very successful program, however, will not result in employment for every participant. Many factors that have nothing to do with the effectiveness of the training program will affect a participant’s employment prospects, such as economic conditions in the community and prior work experience. On the other hand, some program participants would have found jobs even without the assistance of the program. The overall program effect is typically represented as the average effect across all participants and, in that form, depicts the change in the probability of finding a job that was caused

by participation in the program without specifying which particular individuals would or would not have found a job without the program.

Although the main goal of impact evaluation is determining whether the desired effects were produced, this also entails estimating the magnitude of those effects. Stakeholders and other decision makers will want to assess the size of an effect when forming their judgments about a program. If a program is not having the intended effects or the magnitude of those effects is too small relative to expectations, key stakeholders may consider changes in the program, perhaps reviewing the logic of the underlying program theory or assessing whether the program was implemented with fidelity. Findings from impact evaluations that indicate no discernible program effects or negative effects may raise questions about the continuation of the program and the possibility that other approaches may better meet the goals set for the program. On the other hand, when positive effects are found, the discussions often focus on program continuation and possibly even expanding its mission. The potential for influencing these types of high-level decisions underscores the value of impact evaluation.

Additional Impact Questions

Although the main question for impact evaluation is whether the program affected the intended beneficiaries in the ways expected by the program stakeholders, there are other questions that may also be important for an impact evaluation to address. One such question focuses on possible unanticipated consequences of the program. There may be **negative side effects** like those of frequent concern in medical research. For example, a set of impact evaluations known as the Income Maintenance Experiments conducted some years ago focused almost entirely on a potential negative side effect. The program, which offered a guaranteed minimum income for families living in poverty, had several advantages over existing government programs, such as providing additional income to the working poor and ease of administration. However, policymakers were concerned that a guaranteed minimum income might provide a disincentive for participation in the labor force. The last and largest of these impact evaluations, conducted in Seattle and Denver, found that this program reduced adult male work by about 9% and adult female work by roughly 20% (Skidmore, 1985). Whether those reductions may actually be a good thing is debatable—more women, for instance, may have stayed home to take care of young children—but the magnitude of the reduction in labor supply found in these evaluations was important in the subsequent policy debates about how the U.S. government should provide assistance to the working poor.

Another kind of impact evaluation question asks about differential effects: how much variability there is around the average program effect and what factors are associated with that variability. One such question that is often of interest relates to possible differential effects for different subpopulations among the intended beneficiaries. For example, a program to aid the homeless may serve a number of distinct subgroups that will not necessarily react the same way to the services the program provides. It may therefore be important for an impact evaluation to disaggregate the overall average program effect to reveal any differential effects on, say, adult men suffering from mental illness or substance abuse, female-headed families fleeing domestic violence, and LGBTQ youth who have been displaced from their homes. Identification of such differential effects informs program stakeholders about the subgroups that most benefit from the program and

those that benefit the least, or perhaps are even made worse off. That information, of course, has important implications for improving program services, or perhaps developing new services or programs for those not well served by current practice.

Another common concern is differential effects associated with the amount and quality of the services different participants receive from the program, a key aspect of how well the program is implemented. Investigating this source of differential effects is commonly referred to as **dose-response analysis**. Usually evaluators and program personnel expect larger doses of the program to produce larger effects, at least up to some limit. Parenting programs for couples prior to the birth of their child or shortly thereafter, for instance, often involve a curriculum delivered over a certain number of sessions. Dosage in this case relates to the number of meetings attended by either one or both parents and perhaps to how well those sessions inform and engage them. If there is variation in these features but no dose-response relationship is evident, it raises questions about whether the program has any effects or is even needed. When a dose-response relationship is demonstrated, it not only indicates that the program likely makes a difference but yields insight about the level of service needed to produce at least minimal benefits for the participants.

More generally, it can be important for an impact evaluation to explore the influence of variation in how well the program is implemented as a total package. In [Chapter 4](#), we introduced the concept of implementation fidelity, defined as the extent to which a program is implemented as intended by the program designers. Although programs may strive for a high level of fidelity to the program plan, in practice, implementation often varies across program sites and across time in any given site. Assessing fidelity and the associated description of what was actually implemented are essential to defining the program configuration that produced whatever effects are found in the impact evaluation. This information is critical for replicating an effective program and for maintaining the effectiveness of the given program. Moreover, information on fidelity of implementation can aid interpretation of the impact findings. If program impacts are less than anticipated or no discernible impacts are found, implementation data can help establish whether that is plausibly the result of poor implementation of

what otherwise might be a good program. Alternatively, adequate implementation fidelity with no discernible effects suggests that the action theory that guides the program's approach to the problem addressed may not be valid, which we previously referred to as theory failure.

For these reasons, collecting and analyzing data on program implementation is often a component of impact evaluations, and for large federally funded impact evaluations, it is generally expected. Assessing implementation fidelity, however, requires that the program developers, key stakeholders, and evaluators agree on the essential elements of the program plan so that fidelity to that plan can be measured. That, in turn, requires a relatively well developed program theory as the basis for the program's action plan, as discussed in [Chapter 3](#). When that has been adequately formulated for a program, assessing implementation fidelity is relatively straightforward. Indeed, some programs have written manuals or protocols that describe how it is to be implemented. That is not necessarily the case for many ongoing programs, however, and it may require a separate effort by the evaluator to work with the relevant stakeholders to make explicit their tacit understanding of how the program is supposed to be implemented.

In [Exhibit 6-A](#), we provide a list of the objectives and types of questions that commonly shape an impact evaluation. Other than determination of whether the intended effect was produced and estimation of its magnitude, the other questions may or may not be pertinent for any particular impact evaluation. However, they should all be carefully considered when an impact evaluation is being planned. Addressing these additional questions can provide information that will help elaborate a full picture of the nature and extent of the effects of the program and help explain why better or worse effects occurred. Furthermore, for the evidence generated by impact evaluations to guide the development of even more effective programs than those evaluated, it is essential for it to go beyond indications of what works or does not work to address questions of what works for whom under what circumstances and why.

Exhibit 6-A Common Questions Addressed in Impact Evaluations

Impact Evaluation Objectives	Questions to Be Answered
Average impact	What is the average difference in the desired outcomes that is attributable to the influence of the program? Are there unintended beneficial or adverse effects of the program?
Subpopulation average impacts	What is the average program impact on relevant outcomes for different important subpopulations?
Dosage effects	Are more program services and/or higher quality services associated with better outcomes?
Fidelity of implementation	How closely does program implementation match the program plan for the intended implementation? How much does the fidelity of program implementation vary across time, sites, or individuals? Is greater fidelity of implementation associated with larger program effects?

When Is an Impact Evaluation Appropriate?

In principle, impact evaluation is appropriate for any program whose mission includes bringing about change in some set of identifiable outcomes for a defined population or circumstance and for which there is sufficient uncertainty about whether that is being accomplished to justify a need for evidence. As discussed in [Chapter 1](#), whether a program produces its intended effects may be uncertain even when key stakeholders are convinced by their own experience that it is effective. The need for credible evidence, if not already in hand, may be for purposes of accountability, especially for publicly funded programs, but may also be desired to guide program improvement. In practice, most social programs have not been evaluated for impact, and their administrators, sponsors, and advocates have not initiated impact evaluations or been required to do so. Nonetheless, there are various points in the life course of a social program when impact evaluation is especially apt.

At the stage of policy formulation, it is often wise for policymakers to commission a pilot demonstration program with an impact evaluation to determine whether a proposed program can actually have the intended effects. This type of impact evaluation is sometimes referred to as an efficacy trial and is designed to provide proof of concept. That is, it investigates whether the program *can* produce the intended effects under favorable circumstances, for example, with the program developers involved, a small-scale implementation, and a selected, especially appropriate group of recipients. It does not establish that when implemented at scale in routine practice, it *will* have the intended effects. However, if the program is not successful in a small-scale pilot trial, it is very unlikely to be successful if implemented on a broader scale.

Another point in the development of a program that can be especially appropriate for an impact evaluation is when it is being rolled out for the first time. When a new program is authorized, it often cannot be implemented at the ultimately desired scale all at once. It may then be phased in with implementation beginning in a limited number of sites. Impact evaluation at that point can reveal whether the program is producing

the expected effects before it is extended to broader coverage in later phases. A similar situation occurs when the sponsors of innovative programs, such as private foundations, implement programs on a limited scale and conduct impact evaluations with a view to promoting adoption of the program by legislative action or through government agencies if the desired effects can be demonstrated. However, new program implementations can be problematic in ways that should raise concerns for evaluators. In the early stages of a new program, impact evaluation may be premature. For programs of any complexity, it takes time to achieve full implementation—staff must be recruited and trained, operational procedures and policies must be instituted, and the intended beneficiaries need to be reached and engaged. An impact evaluation during the rollout of a program should be considered only if implementation fidelity can be assessed concurrently and, further, when there is a reasonable expectation that implementation fidelity can be achieved rather rapidly or the evaluation will continue through sufficient implementation cycles for fidelity issues to be addressed.

There are also circumstances when impact evaluation is especially appropriate for ongoing programs. For example, there may be a time when a program is modified and refined to enhance its effectiveness, accommodate revised program goals, or reduce costs. When the changes are major, the modified program may warrant impact assessment because it is, at least to some extent, a new and different program. Impact evaluation at that point can ascertain whether the modified program has the intended effects and provide input for further refinements to boost effectiveness.

There may also be good reason to subject a stable, established program to impact assessment. For example, the high costs of certain medical treatments make it essential to continually evaluate their effects and compare them with other means of dealing with the same problem. Long-established programs may be evaluated because of sunset legislation requiring evidence of effectiveness for funding to be renewed, to satisfy demands for accountability, or to defend against attack by critics. An impact evaluation can thus be appropriate at different stages of a program's development, from a demonstration pilot to an ongoing mature program.

At whatever point in a program's development an impact evaluation is undertaken, however, consideration should be given to the scope of information that will be needed to support interpretation of the findings. Input from two of the domains of evaluation discussed in prior chapters stand out in this regard: assessment of program theory and evaluation of program process and implementation. An examination of the program theory allows the evaluator to determine if the program's objectives are sufficiently well articulated and the relationships between activities and outcomes are sufficiently plausible to make it reasonable to expect the program to produce the intended effects. Moreover, the presumption that the activities specified in the program theory are actually implemented with sufficient fidelity, consistency, and quality to yield the expected effects should be grounded empirically as part of the impact evaluation rather than simply assumed. It would be a waste of time, effort, and resources to evaluate the impact of a program that lacks a plausible theory of action for attaining socially significant outcomes or has not been adequately implemented.

It is also important to recognize that the more rigorous forms of impact evaluation involve significant technical and managerial challenges. The intended beneficiaries of social programs are often difficult to reach or may be reluctant to provide outcome and follow-up data. As described in later chapters, impact designs can be demanding in both their technical and practical aspects. In addition, impact evaluation often faces political challenges. Without sacrificing their independence and while contending with inherent pressures to produce timely and valid results, the evaluators must cultivate the cooperation of program staff and participants who may feel threatened by evaluation. Before undertaking an impact evaluation, therefore, evaluators and those sponsoring the evaluation should carefully assess whether it is sufficiently justified by the program circumstances, available resources, and the need for information. Program stakeholders who ask for impact evaluation may not appreciate the prerequisite conditions and resources necessary to accomplish it in a credible manner.

This realistic perspective is not intended to discourage impact evaluation under appropriate circumstances. It is an essential endeavor for answering what is usually the most policy relevant question about a program: Does it

work? If the decision is made to conduct an impact evaluation, the most significant design and planning challenge the evaluator must deal with is how to determine what would have occurred in the absence of the program as a benchmark for assessing the difference in outcomes attributable to the program. This challenge is both distinctive and central to impact evaluation, and we turn to it next.

What Would Have Happened Without the Program?

To isolate the effects of a social program, evaluators conducting impact evaluations need to both measure the outcomes for the individuals exposed to the program and find a credible way to estimate the outcomes that would have occurred in the absence of the program, that is, the outcomes for those same participants at the same time had they not been exposed to the program. The latter—the outcomes in the absence of the program—is not something that can be directly observed or measured. If participants are exposed to the program, we cannot then also know the outcomes they would have experienced had they not been exposed. That part is contrary to the reality that they did, in fact, experience the program. Outcomes in the absence of the program are referred to as the **counterfactual** (contrary to fact), and estimating the counterfactual presents one of the greatest challenges for impact evaluations.

In some physical and laboratory sciences, the counterfactual can be established as the status of an object or research subject prior to applying a hypothesized causal agent, such as heat or a virus. That approach assumes that, in the absence of the intervention, there will be no change in that object or research subject prior to the time when the outcomes are measured. Alternatively, the properties of that object or subject may be so well known that whatever change will occur over that interval is highly predictable, so the researcher can be confident of that prediction as an accurate estimate of the counterfactual. In laboratory contexts, the researcher may control the environment to eliminate other influences that could affect the outcome and thus strengthen the assumption that the counterfactual can be estimated from the initial status of the object or research subject.

In contrast to these situations of predictable outcomes absent the intervention of interest, the excitement and the challenge of evaluation are that the work is performed in the rough-and-tumble world of everyday life. It is extremely rare that evaluators can confidently assume that the intended

beneficiaries of a social program would not have changed in some way that affected their outcomes in the absence of the program. Both through normal growth and human development, and as a result of their own agency and the external environment in which they live, change of a rather unpredictable sort is routine and commonplace for humans. Nor do evaluators have the possibility of controlling the environment in ways that prevent any change from occurring that is extraneous to the intervention being evaluated.

An example may help clarify this point with a little levity. Smith and Pell (2003) ask why there are no rigorous evaluations of the effectiveness of parachutes for “preventing major trauma related to gravitational challenge.” They suggest that studies be conducted, which would truly be “impact” evaluations, that compare health outcomes for individuals who jump out of airplanes with parachutes and those who jump without parachutes. The latter condition is intended to provide an estimate of the counterfactual: the outcome in the absence of the intervention, use of a parachute. The absurdity of this satire, but also its lesson for us, is that we know what the counterfactual outcome is: near certain death. When the outcome absent intervention is totally predictable, no fancy evaluation designs are needed to obtain a counterfactual benchmark against which the program effect can be measured. It is the rarity of that situation that challenges the evaluator to find a way to empirically estimate the counterfactual outcomes when asked to determine the effects of a social program.

This example, although rather extreme, gives us a starting point for how to think about devising a sound counterfactual condition for an impact evaluation. Measures of participants’ status on the target outcomes and other factors prior to program exposure might yield a workable counterfactual, but only if they provide sufficient information to accurately predict the outcomes that would be found later if those participants were instead not exposed to the program. Though relatively rare, there are circumstances in which this may be the case, for instance, when the outcomes at issue relate to stable conditions unlikely to change on their own. Consider a lead paint abatement program in public housing. There is little that would cause lead paint to disappear absent a program to remove it, so the initial conditions may be a valid counterfactual. If the prevalence of lead poisoning among children living in the public housing is the target

outcome, however, the evaluator must be alert to other sources of lead poisoning that might arise in the interim. As we know from Flint, Michigan, for instance, changes in the water supply could create a new source of lead exposure for children.

In the more common situation in which the counterfactual outcomes are uncertain, preintervention conditions will not provide an accurate estimate. A reasonable alternative would be to consider using the outcomes for a group of individuals who did not participate in the program as the counterfactual benchmark for determining the effects of the program for those who did participate. For this approach to provide a sound counterfactual estimate, however, the individuals who do not participate in the program would have to be similar to those who do on any characteristic related to the later outcomes. That is, the two groups must be comparable in ways that would yield the same outcomes for both in the absence of exposure to the program. That can be a difficult standard to meet. There are typically multiple, mostly unknown reasons why some individuals participate in a program and others do not, any of which might influence the postintervention outcomes. Because participation in most social programs is voluntary, for instance, those who choose to participate may have more motivation to improve their outcomes or the presence of supportive family members who can support their efforts. Even without the program, such individuals might be expected to have different outcomes than those who chose not to participate. For programs, such as job training programs or college access interventions, program staff may select individuals on the basis of some eligibility criteria, creating potentially problematic differences between those selected and those not selected. Even when there is not such readily apparent deliberate selection into program participation, there are generally inherent natural selection processes, such as differential opportunity or capacity, geographical proximity, and the like, that have acted to sort individuals into program participants and nonparticipants.

These selection processes can easily result in differences between program participants and nonparticipants that, in turn, can lead to different outcomes unrelated to actual program effects. Because of the potential for such differences, known as **selection bias**, evaluators cannot confidently assume that the outcomes for those who did not participate in a program would be a

valid estimate of the counterfactual condition for those who did participate. Selection bias can represent initial differences between participants and nonparticipants that are directly related to the outcomes of interest, or differences associated with the reaction to the program, such as motivation, social support, or engagement. These two sources of selection bias, initial differences and differences in response to treatment, are highly salient concerns in nearly every impact evaluation, making selection bias the most common type of bias that must be dealt with in impact evaluations.

The distinctive difficulty of conducting impact evaluations should now be apparent. The outcomes of interest for most programs are factors that often change over time for the intended beneficiaries, whether they participate in a program or not. Moreover, selection bias may cause differences to appear in the outcomes of individuals who participate in a program relative to those who do not participate that may look like program effects but, in fact, are not. Yet to determine the effects of a social program, impact evaluations must provide plausible and credible answers to the question: How much better off are the program participants than they would have been had they not participated in the program? Before describing the particular techniques and procedures evaluators can use to deal with this situation, we lay out the overall logic for tackling the challenges of impact evaluation.

The Logic of Impact Evaluation: The Potential Outcomes Framework

As noted, impact evaluation requires a credible counterfactual that allows evaluators to estimate the outcomes that would have occurred in the absence of the program. A framework for impact evaluation that has been developed and refined in recent years aids our understanding of that logic and helps us identify the assumptions needed to regard a program effect found in an impact evaluation as sound and convincing. This framework is known as the **potential outcomes** framework. It was originally proffered by a statistician, Donald Rubin, who has also contributed greatly to its refinement and application to program evaluations (Holland, 1986). The potential outcomes framework guides evaluators' efforts to determine the effects of known causes, which must be distinguished from attempts to determine the causes of known effects. The social programs, policies, or interventions of interest for impact evaluation are the known causes in this formulation, and the job of the evaluator is to determine their effects on the targeted outcomes. Attempting to determine the causes of known effects, by contrast, requires a backward look from outcomes to identify what produced them. That is the kind of work epidemiologists do when, for instance, they try to determine what caused an outbreak of a certain disease.

For any individual, we expect that the experience of being exposed to a program will cause better, or at least different, outcomes to occur than with no exposure. In other words, any such individual has two potential outcomes: one that would occur with exposure and another that would occur without exposure. These outcomes can be the same or different. If they are the same, the program has no effect for that individual; if they are different, the program does have an effect, one defined by that difference. The potential outcomes for different individuals in relation to any given program can be different, and we generally assume they are. The overall effect of the program on the individuals exposed to it is thus determined by the mix of potential outcomes for that group of individuals.

How this works can be illustrated with a simple example. Assume for the moment that the outcome of interest is dichotomous: success or failure. Many outcomes take this form. A student in an alternative high school might graduate or not. A participant in job training might or might not be employed afterward. A youthful offender in a juvenile justice rehabilitation program will or will not reoffend. For such dichotomous outcomes, each member of the target population has one of four possible combinations of potential outcomes, as shown in [Table 6-1](#).

Table 6-1

TABLE 6-1 Possible Potential Outcomes

		If Exposed to Program	
		Success	Failure
If Not Exposed to Program	Success	A Bulletproof	C Backfire
	Failure	B Bull's-eye	D Out of range

Individuals whose potential outcomes are characterized by Cell A achieve a successful outcome whether they are exposed to the program or not. We might think of these individuals as bulletproof: they succeed with or without the program. Individuals with potential outcomes characterized by Cell B succeed if they are exposed to the program, but fail if not exposed to the program. These individuals represent program bull's-eyes; exposure to the program changes their outcomes from failure to success. The individuals in Cell D fail whether they are exposed to the program or not. We might say that these individuals are out of range of the program: for them, exposure to the program is not sufficient to change failure into success, though an alternative program may be able to do that.

The individuals in Cell C have positive outcomes if not exposed to the program but fail if they are exposed to the program. These are individuals for whom the program has backfired. This may appear to be an unlikely combination of potential outcomes, but consider a substance abuse prevention program that aims to dissuade youth who have not yet used drugs from doing so. Some of those youth wouldn't use drugs anyway; they

do not need a prevention program to have a successful outcome. Suppose now that the prevention program exposes these youth to information about some drugs and their effects that they did not know about, and, the adolescent brain being what it is, that tempts them to try the drugs rather than dissuading them. For them the program has backfired. An example of a program for which the backfires equal or exceed the bull's-eyes, though the reason is not clear, is D.A.R.E. (Drug Abuse Resistance Education), a popular school prevention program that some impact evaluations show actually increased drug use among adolescents and, when effects from many studies are combined, shows no effect (West & O'Neal, 2004).

The important takeaway from [Table 6-1](#) is that the direction and magnitude of program effects for a target population depend on the proportions of individuals with different combinations of potential outcomes. When the proportion of individuals in Cell B (bull's-eyes) exceeds that in Cell C (backfires), the program has an overall positive effect, albeit not necessarily for every participant. However, a relatively large proportion of the target population in Cell A or Cell D can overwhelm the differences in Cells B and C and attenuate the overall program effect toward zero. [Table 6-2](#) illustrates the interplay between the proportions of the target population in the difference potential outcome cells on the overall program effect. For these hypothetical examples, we present the program effect as the ratio of the proportion of successes to the proportion of failures when exposed to the program divided by the ratio of successes to failures without program exposure (an index called the odds ratio). When this ratio is greater than 1, there is a positive average program effect. When it equals 1, there is no effect, and when it is less than 1, the average program effect is negative.

Table 6-2

TABLE 6-2 Hypothetical Average Program Effects for Three Scenarios Using the Potential Outcomes Framework With a Target Population of 250 Individuals

	Example 1: More Bull's-Eyes		Example 2: More Backfires		Example 3: More Bulletproofs	
	Successes With Program	Successes Without Program	Successes With Program	Successes Without Program	Successes With Program	Successes Without Program
Cell A: bulletproof	50	50	50	50	100	100
Cell B: bull's-eye	100	0	50	0	50	0
Cell C: backfire	0	50	0	100	0	50
Cell D: out of range	0	0	0	0	0	0
Total successes	150	100	100	150	150	150
Total failures	$250 - 150 = 100$	$250 - 100 = 150$	$250 - 100 = 150$	$250 - 150 = 100$	$250 - 150 = 100$	$250 - 150 = 100$
Odds of success	$150/100 = 1.50$	$100/150 = .667$	$100/150 = .667$	$150/100 = 1.5$	$150/100 = 1.5$	$150/100 = 1.5$
Ratio of odds of success with and without program	$1.50/.667 = 2.25$ (positive average effect)		$.667/1.50 = 0.45$ (negative average effect)		$1.50/1.50 = 1.00$ (null average effect)	

The first example in [Table 6-2](#), in which the potential outcomes for the target population include more bull's-eyes than backfires, shows an overall average positive program effect as indicated by greater odds of success if exposed to the program than if not exposed. Note that if there were no backfires, the average positive effect would be driven entirely by the bull's-eyes and would be even larger. Furthermore, if the proportion of bulletproof cases (adding equal successes both with and without the program) were increased, or the proportion of out-of-range cases (adding equal failures both with and without the program), the average program effect would still be positive but smaller. Similarly, in the second example the proportion of backfires exceeds that of bull's-eyes, producing a negative average program effect (odds ratio < 1), which would be even more negative if there were no bull's-eyes and smaller, but still negative, if the proportion of bulletproof or out-of-range cases were larger.

Near the beginning of this chapter, we pointed out that cause-and-effect relationships for programs were probabilistic. When we speak of a program causing an effect on some outcome, we mean that it increases the probability of that outcome appearing among the members of the target population. The potential outcomes framework allows us to put a finer point on one source of the probabilistic nature of program effects. First, the

increased probability of an outcome produced by an effective program is a relativistic concept: it is the difference between the likelihood of that outcome in a target population with program exposure relative to its likelihood without such exposure. This is illustrated by the successful potential outcomes without program exposure that are shown in the examples in [Table 6-2](#). Success is possible with and without program exposure, the effect of the program is the difference in the probabilities of those potential outcomes.

Second, the direction and magnitude of the program effect is a function of the mix of different patterns of potential outcomes present in the target population. That too can be viewed as probabilistic (e.g., the likelihood that there are fewer or more bull's-eye patterns of potential outcomes for a given program in the target population along with all the other potential outcome patterns that are not so favorable for the program). With outcomes that involve varying degrees of success or failure, such as income, academic achievement, and obesity, there are even more patterns of potential outcomes in the mix for a target population than in the examples used in [Table 6-2](#), and thus a more complex set of probabilities associated with the proportions in that mix. There are other probabilistic aspects of the estimates of program effects associated with the methods used to generate those estimates that will warrant attention in later chapters. However, the potential outcomes framework reveals that the probabilistic nature of program effects is inherent in concept of a program effect under conditions of different potential responses to program exposure among the target population.

The Fundamental Problem of Causal Inference: Unavoidable Missing Data

The potential outcomes framework provides evaluators with a conceptual framework for understanding the nature of program effects and the challenges associated with assessing them. In particular, it highlights the role in the overall program effect for a target population of the potential outcomes with and without program exposure for each individual or unit in that population. For each such unit it is not possible to simultaneously observe the outcomes with and without program exposure. This is known as “the **fundamental problem of causal inference**,” and it means that when the outcomes for those exposed to the program are observed, their potential outcomes without program exposure must somehow be inferred in order to determine the program effect. The potential outcomes without program exposure, of course, are the counterfactual outcomes discussed earlier in this chapter that are fundamental to the definition of a program effect.

The dilemma presented by this situation can be characterized as a missing data problem. When the impact evaluator collects data on the outcomes for program participants, the data on the potential outcomes that represent the counterfactuals for those same participants at that same time are automatically and unavoidably unavailable. In order to calculate a program effect, something must be done to find a value for these missing data points. Whatever is done, it will not be direct measurement of the “real” potential outcomes absent treatment, but an estimate of some sort. The difference between the observed outcomes with the program and the estimates of the counterfactual outcomes without the program that constitutes the program effect will thus also be an estimate, and its accuracy will depend in large part on how good the estimation of the counterfactual is.

As noted, potential counterfactual outcomes reside at the level of the individuals in the program’s target population. It is very rare to find a situation in which convincing individual-level counterfactual outcome estimates can be made in evaluation research. It would be necessary for the evaluator to make highly accurate predictions of the outcomes without the

program for each individual, such as those expected in the example of jumping out of airplanes without parachutes. Such predictions are not possible for the kind of counterfactual outcomes at issue for most social programs. Alternatively, preintervention baseline measures of relevant outcomes for each individual could provide good individual-level counterfactual estimates, but only if it is safe to assume that no change would occur before the time of outcome measurement, or that whatever change will occur is completely predictable. Stable physical situations, such as the lead paint in low-income housing (with houses as the relevant individual unit) in our previous example, may provide such circumstances, but they too are rare in impact evaluation for social programs.

Instead of individual-level counterfactual estimates, evaluators most often find it necessary to rely on group-level estimates. A common way of doing this is by constructing or identifying a group of individuals who did not participate in the program being evaluated whose outcomes of interest can be averaged to use as a counterfactual estimate for the average of the group that did participate in the program. The difference between those averages then becomes the estimate of the overall average program effect. Depending on the similarity of the groups and the potential for selection bias we discussed earlier, this approach can yield good estimates of overall average program effects, and generally also for average program effects for some subgroups. However, it does not produce a counterfactual estimate for each individual in the program group.

The chapters that follow this one provide an overview of the various research designs impact evaluators can use to develop valid estimates of program effects, with the way the counterfactual outcomes are estimated as the main feature distinguishing the different designs. [Chapter 7](#) describes what are generally called comparison group designs: those that do not strictly control who receives access to the program and who does not. [Chapter 8](#) then describes what are generally called controlled designs, in which there are strict controls on access to the program.

The Validity of Program Effect Estimates

As we trust this chapter has made clear, impact evaluation is an especially challenging endeavor. The program effects it attempts to estimate are themselves quite problematic because of the need to find data to represent the inherently unobservable counterfactual potential outcomes. Along with the efforts needed to adequately measure relevant outcomes of those with exposure to the program after that exposure occurs, the practical aspects of impact evaluation also demand that the evaluator come up with convincing estimates of those counterfactual outcomes. Under these circumstances, an overarching concern for all of impact evaluation is the validity of the resulting program effect estimates.

The main types of validity for research on causal relationships such as those between a program and its target outcomes are well defined and relevant to every impact evaluation. We first note that although we have referred frequently to program effect estimates for the target population of a program, impact evaluation is not typically done for the entire target population or even for the entire subset of that population that is actually exposed to the program. As a practical matter, impact evaluation is usually done with a subset of the individuals who are exposed to the program, that is, with a selected sample of the target population, referred to as the participant study sample.

A central concern for impact evaluation is the **internal validity** of the program effect estimates. *Internal validity* refers to the validity or accuracy of an effect estimate for the specific participant study sample used in the impact evaluation. In theory, an internally valid effect estimate reflects the actual effect that would be found if both values of the potential outcomes could be known for the participant study sample. In practice, given the impossibility of that, internal validity is high when complete outcome data for the participant study sample and accurate and complete measures of the relevant counterfactual outcomes are used to compute the program effect estimates. The validity of the resulting effect estimates, however, will be limited to those in the particular study sample of participants.

Every impact evaluation should aspire to have high internal validity. Without that, the conclusions reached about the direction and magnitude of the program effects may simply be wrong and, therefore, quite misleading for program stakeholders who want to know if the program has the intended impact on participants. Nonetheless, if the participant study sample for an impact evaluation is not the entire target population, there is another validity issue to consider, known as **external validity**.

External validity is the extent to which the program effect estimates derived from the study sample accurately characterize the program effect for the full target population, which is often called *generalizability* of the program effect. The study sample used in the evaluation may be quite similar to the target population with regard to the characteristics that influence the outcomes of interest, especially with regard to the factors related to the outcomes prior to exposure to the program and the way individuals in the target population respond to the program. In that case, external validity is high: the program effects for the full target population that were not directly estimated should be similar to, or generalizable from those found for the evaluation sample. But if the evaluation sample is different in ways that relate to the relevant outcomes, then the program effects found for that sample, whatever their internal validity, may also be different from those that occur for the full target population. Under those circumstances, external validity would be low. The best way to ensure external validity is to draw a representative study sample from the target population, for example, a probability sample from a well-defined population, but that often proves impractical in many evaluation circumstances. When we describe the major research designs used in impact evaluation in the two chapters that follow, we will frequently describe their implications for internal validity—the extent to which the program’s effect estimate for the subset of the target population used in the evaluation is accurate—and external validity—the extent to which an evaluation program’s effect estimate accurately characterizes the program effect for the entire target population.

Summary

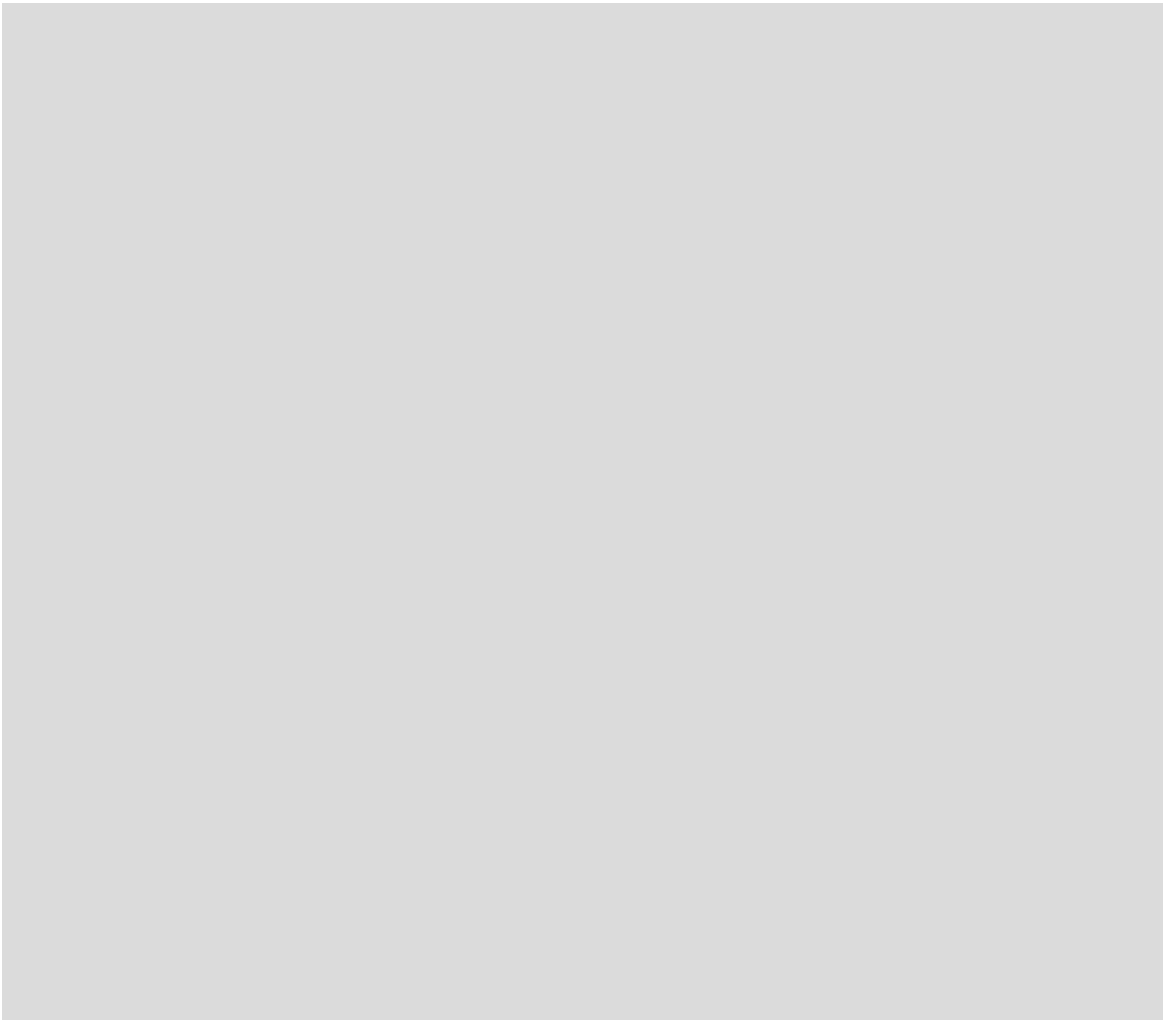
- Impact evaluation addresses a high-priority question: whether the program brings about the intended beneficial changes in the target population. Because of its

potential to influence policy and high-level program decisions, it is one of the most important forms of evaluation.

- Identifying and measuring the program effects is a matter of demonstrating that the program has caused change in the outcomes for the participants that would not otherwise have occurred. Impact evaluation thus fundamentally involves cause-and-effect relationships in which exposure to the program is expected to cause a change in the probability of desirable outcomes.
- Although the main question for impact evaluation is whether the program had the intended effects, other issues may also be relevant, for example, possible unanticipated positive or negative effects, differential effects for different subpopulations, and varying effects related to the amount and quality of the services or fidelity to the program design.
- Impact evaluation is appropriate in concept for any program intended to bring about change and for which there is uncertainty about whether that is being accomplished. It may be especially appropriate for early pilot and demonstration programs, when a new program is first rolled out, and when an ongoing program is modified in ways that might affect the outcomes.
- To isolate the effects of a social program, impact evaluators must measure the outcomes for individuals exposed to the program and compare them with estimates of the outcomes that would have occurred for those individuals in the absence of the program, which is called the counterfactual.
- The counterfactual outcomes necessary to assessing program effects cannot be observed but may be estimated in various ways depending on the circumstances. Possible approaches include using information that allows confident prediction, initial baseline outcome values if they can be assumed stable or can accurately predict later outcomes absent intervention, and outcomes for untreated comparison groups sufficiently similar to program participants.
- The potential outcomes framework provides the conceptual underpinnings for impact evaluation. Each individual in a program's target population has one potential outcome that will appear with program exposure and another that will appear if there is no exposure. The difference between them is the program effect for that individual, and the overall program effect is a function of the mix of potential outcome patterns in the target population and the probability with which each pattern occurs.
- Potential outcomes with and without program exposure cannot be simultaneously observed (known as the fundamental problem of causal inference). When outcomes are measured for program participants, the unobservable counterfactual potential outcome absent the program can be viewed as missing data that must be handled with a convincing estimation procedure. The major approaches for that are reviewed in [Chapters 7](#) and [8](#).
- An overarching concern for all impact evaluation is the validity of the resulting program effect estimates. An effect estimate has internal validity when it is an accurate representation of the actual effect for the program participants for which it is estimated. That effect estimate has external validity if it also generalizes to the full target population, even though not all of them participated in the evaluation.

Key Concepts

Counterfactual 147
Dose-response analysis 143
External validity 154
Fundamental problem of causal inference 152
Impact evaluation 142
Internal validity 154
Negative side effect 143
Potential outcomes 149
Program effect 141
Program impact 141
Selection bias 148



Critical Thinking/Discussion Questions

1. Although impact evaluations are necessary to assess a program's effects on its target outcomes, most programs are not evaluated. Identify three times in the life course of a social program when an impact evaluation might be appropriate, and explain how the impact evaluation could be used at those times.
2. Outcomes in the absence of the program are referred to as the counterfactual. Estimating the counterfactual presents one of the greatest challenges for impact evaluations. Explain why this is so challenging.
3. Explain what is meant by the "fundamental problem of causal inference" and why it can be viewed as an unavoidable missing data problem.

Application Exercises

1. Using the potential outcomes framework, propose a social intervention with its target outcomes. Then create a table showing the potential outcomes for participants in that program (like [Table 6-1](#)). Explain the situation represented in each of the possible outcomes represented in that table.
2. With the same social intervention you used above, expand on the average program outcome that might result from different mixes of the potential outcomes you identified above (like [Table 6-1](#)). On the basis of your understanding of the social intervention, which average outcome do you think will be most likely and why?

Chapter 7 Impact Evaluation Comparison Group Designs

Bias in Estimation of Program Effects

Selection Bias

Other Sources of Bias

Secular Trends

Interfering Events

Maturation

Regression to the Mean

Potential Advantages of Comparison Group Designs

Comparison Group Designs for Impact Evaluation

Naive Program Effect Estimates

Covariate-Adjusted, Regression-Based Estimates of Program Effects

Multivariate Regression Techniques

Program Effect Estimates From Matched Comparison Groups

Choosing Variables to Match

Exact Matching and Propensity Score Matching

Interrupted Time Series Designs for Estimating Program Effects

Cohort Designs

Difference-in-Differences Designs

Comparative Interrupted Time Series Designs

Fixed Effects Designs

Cautions About Quasi-Experiments for Impact Evaluation

Summary

Key Concepts

In this chapter we discuss designs for impact evaluation in which the counterfactual outcomes are estimated from comparison groups that were not exposed to the program. Because comparison groups, as defined in this chapter, are not recruited or constructed in a way that ensures that they will support valid estimates of program effects, designs that rely on them are vulnerable to various sources of bias. After cautions about the ways in which estimates of program effects can be biased in these designs, we describe four types of comparison group designs that are useful in many circumstances in which an impact evaluation is required. The advantage of these designs is that they are less intrusive for

the programs being evaluated than a more controlled design and thus are often more feasible to implement for practical reasons. For each of these four types of comparison group designs, we identify the defining characteristics, illustrate applications, and review potential sources of bias. In conclusion, we remind the reader that better controlled designs are preferable when feasible, and that comparison group designs have limitations that must be acknowledged and overcome whenever possible.

As we described in [Chapter 6](#), impact evaluations that assess the effects of programs on their target outcomes are prized for their potential to influence policy and high-level program decisions. Also noted was the inherent comparative logic of impact evaluations: What is meant by a **program effect** or **program impact** is the difference between the outcome for members of the target population exposed to the program and the outcome that would have occurred, all else equal, if the same individuals were not exposed to the program (the counterfactual condition). Measuring outcomes for those who participated in a program is generally rather straightforward. Coming up with valid estimates of outcomes for the hypothetical counterfactual condition, however, is a major challenge for impact evaluation. Some approaches to this challenge are especially attractive to evaluators because of the relative ease with which they can be implemented.

One of these approaches involves the use of a **comparison group** drawn from some pool of program nonparticipants. In this impact evaluation design, outcomes measured for individuals exposed to the program (the **intervention group** or **program group**) are compared with those for more or less similar individuals who were not exposed to the program for whatever reason. Those reasons may involve individual choice; administrative criteria or staff discretion for eligibility, priority, or capacity for enrollment; lack of access to the program; or other such circumstances that yield a group of nonparticipants who can be recruited for the evaluation.

Another approach is to focus on change from before program exposure to after for a group of individuals exposed to the program. In this approach, the evaluator must identify individuals who are expected to participate in the program, but have not yet begun, and arrange for measurement of the target outcome prior to their participation as well as afterward. The status on an outcome measure before program exposure is used to estimate the

counterfactual condition, and the before-after comparison then becomes the basis for estimating program effects.

What these approaches and their variants have in common is that the evaluation design does not require that access to the program be strictly controlled, such as by a lottery to determine who can participate and who cannot. The strongest impact evaluation designs rely on strict controls on which members of the target population are given the opportunity to participate in the program and which are not offered that opportunity. This control over the conditions used to estimate the counterfactual outcomes strengthens such designs. Controlled designs of this kind are described next in [Chapter 8](#). The designs discussed in this chapter do not involve such control but, rather, take advantage of naturally occurring differences in program exposure, whether between groups, over time, or both. These are often referred to as *comparison group* designs in contrast to *control group* designs, and we will use that terminology here.

Executed well under favorable circumstances, comparison group designs can provide valid estimates of counterfactual outcomes and, therefore, valid estimates of program effects. However, in comparison with controlled designs, they are more vulnerable to a range of potential biases that can undermine the validity of those estimates. The attractiveness of these designs for impact evaluators due to their relative convenience must therefore be tempered by acknowledgment of those vulnerabilities. Although a major concern of evaluators using any impact evaluation design should be to minimize bias in the estimate of program effects, such efforts are especially important for comparison group designs.

In this chapter, we describe four types of comparison group designs that can be used for impact evaluation:

1. naive estimates of program effects;
2. covariate-adjusted, regression-based estimates of program effects;
3. matching designs, including propensity score matching; and
4. interrupted time series designs including cohort designs, difference-in-differences, comparative interrupted time series, and fixed effects.

First, however, we review the forms of bias that may compromise the results of these designs and the ways researchers can try to guard against them. With that understanding in mind, we then turn to the four designs and describe how they may be used to estimate program effects when a better controlled design is not feasible.

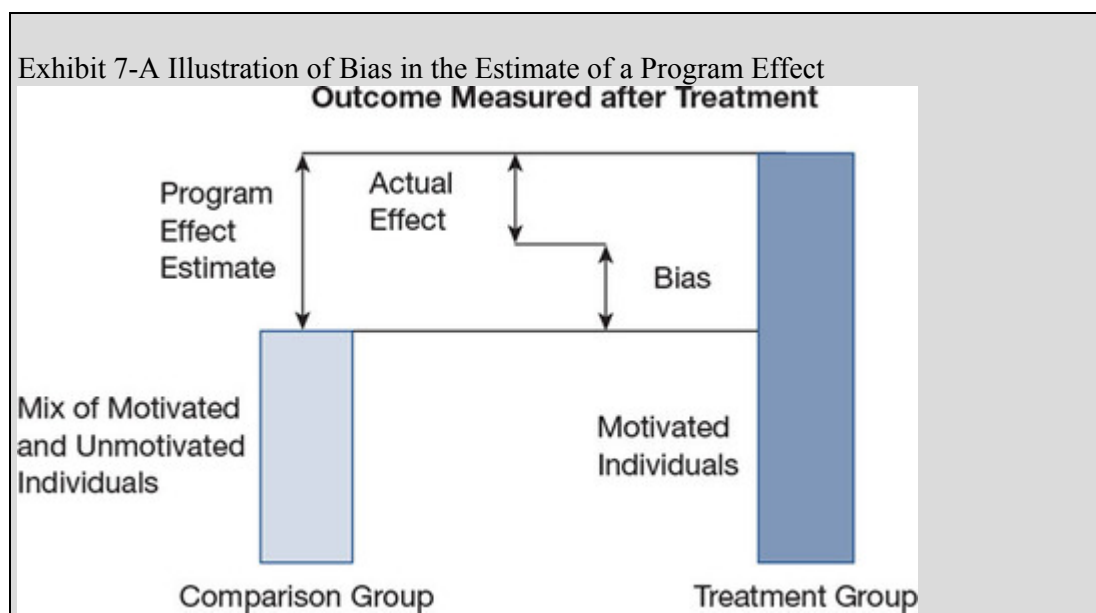
Bias in Estimation of Program Effects

A valid estimate of a program effect results when the difference between the observed outcome with program exposure and the estimate of the counterfactual outcome that would have occurred without exposure are both accurate representations of their respective conditions. Bias is present when either the measurement of the outcome with program exposure or the estimate of the counterfactual outcome departs from the corresponding true value. Unfortunately, the extent of the bias cannot be determined from the data collected for an impact evaluation, leaving some degree of uncertainty about the validity of the effect estimates with even the strongest of these designs.

One potential source of bias comes from the measurement of the outcomes for program participants. This type of bias is relatively easy to avoid by using measures that are valid for what they are supposed to be measuring and responsive to the full range of outcome levels likely to appear among the individuals measured (see [Chapter 5](#) for a discussion of outcome measurement issues in evaluation). A more common source of bias is a research design, or the way it is implemented, that systematically underestimates or overestimates the counterfactual outcomes. Because the actual counterfactual outcome cannot be directly observed, there is no foolproof way to determine whether such bias occurs and, if so, its magnitude. This inherent uncertainty is what makes the potential for bias so problematic in impact evaluations using comparison group approaches. Below we describe some of the most common sources of bias that bedevil impact evaluators.

Selection Bias

If there is no program effect, the outcomes for those exposed to the program and the outcomes for the comparison group used to estimate the counterfactual should be the same. However, if there is some preintervention difference between the program group and the comparison group that is related to the outcome, that difference will cause the outcome to differ in a way that looks like a program effect but, in fact, is only a bias introduced by the initial difference between the groups. This form of bias is known as **selection bias** and was described earlier in [Chapter 6](#). Selection bias gets its name because it arises when some process that is not fully known influences whether individuals enter into the program group or the comparison group with no assurance that this process has selected completely comparable individuals for each group.



For example, an evaluator assessing the impact of a vocational training program could compare employment and wages for those who completed the program with employment and wages for a group of similarly unemployed individuals residing in the same community who did not enroll in the program. In this circumstance, the individuals who enrolled may well have been more motivated to improve their job prospects than those in the comparison group. If motivation itself is related to the likelihood of

obtaining employment, this difference would systematically bias the estimates of the program effects upward. This sort of selection bias is illustrated in [Exhibit 7-A](#) using a graph that illustrates a program effect estimate that includes the bias stemming from the greater motivation of the program participants along with the actual program effect. The issue for the evaluator is that there is no way to disentangle how much of the difference in the outcome is due to bias and how much is the program effect.

Another more subtle form of selection bias can occur when there is a loss of outcome data for members of intervention or comparison groups that have already been formed, a circumstance known as **attrition**. Attrition can occur when members of the study sample cannot be located when outcomes are to be measured, or when they refuse to cooperate in outcome measurement. When attrition occurs, the outcomes of those individuals are no longer a part of the average outcomes for their respective group. If the unobserved outcomes of those no longer in each group differ from those whose outcomes are observed, there will be a corresponding systematic difference in the observed outcomes of those who remain. That difference results from differential attrition and not from an actual program effect and thus represents another form of selection bias. For the vocational training program example above, if the individuals in the comparison group, who have no affiliation with the program, are more difficult to locate or less willing to participate in a follow-up survey of employment status than the program participants, the resulting differential attrition may bias the estimate of the program effect. This would happen, for instance, if outcome data were more likely to be missing for individuals who move frequently and are chronically unemployed, and more outcome data were missing from the comparison group than the program group.

In general, program participants with missing outcome data cannot be assumed to have the same outcome-relevant characteristics as members of the comparison group whose outcome data are missing. It follows that any amount of attrition that is not negligible opens the door to this form of selection bias. Note that *attrition* here refers exclusively to the loss of cases from outcome measurement. Individuals who drop out of the program do not create selection bias if outcome data can be collected for them. Program dropouts or noncompleters degrade program implementation, but not the

validity of the research design for assessing the impact of the program at whatever degree of participation is attained. The evaluator should thus attempt to obtain outcome measures for everyone in the originally configured intervention group whether or not they actually received the full program. Similarly, outcome data should be obtained for everyone in the comparison group even if some ended up receiving the program or another relevant service. If outcome data are obtained for all members of the group, the validity of the design for comparing outcomes for the two groups is retained. What suffers when there is not delivery of complete service to the participant group or a complete absence of such service to the comparison group is the degree of contrast between the two conditions and the meaning of the resulting estimates of program effects. Whatever program effects are found represent the effects of the program as delivered and taken up by the study sample, whether in the designated program or comparison group.

In sum, selection bias is a relevant concern in all situations in which the units that contribute outcome data to a comparison between those intended to receive program services and those not intended to receive such services may differ on characteristics that influence their outcome status aside from those related directly to program participation.

Other Sources of Bias

Apart from selection bias, other factors that may bias the results of an impact evaluation generally have to do with circumstances other than the program that can create a difference in the outcomes of the program and comparison group that mimics a program effect and cannot easily be disentangled from the actual program effect. For example, if one group has experiences other than program participation that affect the outcome that the other group does not have, those experiences will bias the estimate of the program effect. That bias can make the program effect appear larger or smaller than it actually is, depending on whether the other experience has a positive or negative effect on the outcome.

Social programs do not operate under controlled laboratory conditions but, rather, in environments in which ordinary or natural events inevitably influence the outcomes of interest. For example, many persons who recover from acute illnesses do so as a result of natural body defenses rather than externally administered treatment. Impact evaluations of treatments for some pathological condition—influenza, say—must therefore distinguish treatment effects from the effects of such natural processes in order to avoid overestimating the actual effect of the treatment. The situation is similar for social interventions. A program to reduce poverty must consider that some families will become better off economically without outside help. Or, there may be changes in the environment aside from program exposure that can affect outcomes, such as a recession that increases unemployment.

Influences of this kind will bias the results of impact evaluations whenever they affect the outcomes in one of the groups in a comparison group design differently from the other. If both the program and comparison group outcomes are equally affected, no bias will be created when their outcomes are compared. In what follows, we describe some of the kinds of experiences and events that are often of concern in impact evaluations because of their potential to have differential influence in the outcomes in comparison group designs.

Secular Trends

Naturally occurring trends in the community, region, or country, sometimes termed **secular trends**, may produce changes that enhance or mask actual program effects. In a period when birth rates are declining, a program to reduce fertility may appear more effective than it actually is if that trend is not accounted for in the effect estimate. Conversely, an effective program to increase crop yields may appear to have no impact if the estimates of its effects are masked by the influence of poor weather during the growing seasons in the region where the program is implemented that did not occur in the comparison region. Evaluators implementing a comparison group design need to be cognizant of any differential influences of this sort in the communities from which the participant and comparison samples are drawn. Selecting both groups from the same or, if not possible, geographically proximate and otherwise similar communities may reduce the potential for bias from such secular influences.

Interfering Events

Sometimes short-term events can produce changes that distort the estimates of program effect. A power outage that disrupts communications and hampers the delivery of food supplements may interfere with a nutritional program in a way that diminishes program effects below those that would result under more normal circumstances. Similarly, a natural disaster may make it appear that a program to increase community cooperation has been effective, when it is the crisis situation that brought community members together. When such events occur in the program context but not in the comparison context, they produce bias in the estimates of the program effects. For instance, a revenue shortfall during the period when a community development program is implemented may result in fewer services being provided throughout the community than in a comparison period without the program, biasing the program effect estimates toward zero. Aspects of the evaluation itself may be such an **interfering event** if they can influence outcomes and differ for the program and comparison groups. This could occur, for example, if there are data collection activities only for the program group aimed, say, at assessing program

implementation, which include focus groups, surveys, or interviews that trigger a reaction among participants that affects outcomes measured later via self-report.

Maturation

Impact evaluations must often cope with natural maturational and developmental processes that can produce change in a study population independent of program effects, referred to generally as **maturation**. If those changes affect one group in a comparison group design more than the other, they will bias the program effect estimates. Such bias can easily occur in comparisons between groups of different ages. For example, the effects of a second grade reading program on reading gains may be underestimated if it is compared with gains made during first grade by the same children because of the greater natural developmental gains of younger children. Maturation trends can affect older adults as well. A program to improve preventive health practices among elderly adults may show upwardly biased effects on health outcomes in comparison to a group of even older adults because health generally declines with age.

Regression to the Mean

Another potential source of bias is associated with the tendency for more extreme outcomes to naturally drift in a less extreme direction over subsequent time periods. For example, a large spike in the crime rate may stimulate more intense police patrolling in high-crime areas. Such spikes, however, may result from largely chance circumstances unlikely to repeat in the next period, which therefore will more often show a decrease than a similar or greater crime rate. This phenomenon is called **regression to the mean**, which means that the outcomes of interest tend to return to the longer term average that existed before the extreme instance. For the police response to the crime spike, regression to the mean can make their response look effective when it may have had no actual effect on crime. It is not unusual for organizations to adopt new programs or make major modifications in existing ones when the conditions they address take a turn for the worse. When those conditions result from an unusual confluence of

influences, there is likely to be a naturally occurring rebound that can appear as a positive effect of the revised program on the target outcomes.

A more subtle instance of regression to the mean can occur when an attempt is made to match program and comparison group participants on their initial scores on a pretest of the outcome of interest. If the distributions of such scores for the two groups differ, matches will be available only in the area where they overlap. When scores are matched from the high end of one distribution and the low end of the other, those more extreme values are likely to regress to the means of their respective distributions when measured again later to assess the postintervention outcome.

As an example, consider an evaluation of a meditation program aimed at improving the performance of college track athletes who compete in 800-meter races. The evaluator might select runners with similar times in their last event from two different track teams, with those from one team receiving the program and those from the other serving as the comparison group. If the average times for the two teams differ appreciably, the runners with times from their prior event that match will include some runners from the slower team who had an unusually good performance in that event and some from the faster team who had an unusually poor performance in their last event. In later events, runners on the slower team will tend to regress toward the mean for their team, as will those on the faster team. The postintervention difference between the groups will then include a regression-to-the mean artifact that will bias the estimate of the effect of the meditation program on the performance of the participating athletes.

Potential Advantages of Comparison Group Designs

The types of biases that may occur with comparison group designs that we have just described threaten the **internal validity** of the program effect estimates. *Internal validity*, as you may recall from the description in [Chapter 6](#), refers to the ability of a study to detect and produce an unbiased program effect estimate for the units included in the study. Although attempting to minimize threats to internal validity is an essential task when designing comparison group studies, comparison group designs may have advantages for **external validity**. *External validity* refers to the ability to generalize the program effect estimate beyond the units in the study to the broader target population of individuals or units eligible to participate in the program. Comparison group designs do not generally require that participation in the program group be restricted or controlled to meet the inherent requirements of the design itself. This is not always the case with the designs discussed next (in [Chapter 8](#)), such as the randomized control design, that have fewer threats to internal validity but may have to be implemented with only a selected subset of the program population (e.g., individuals willing to volunteer to participate in random assignment).

Comparison group designs, by contrast, often include all the actual program participants in the program group or, for multisite programs, perhaps all those participating in the program at selected sites. In this regard, the program group for which effects on the target outcomes are being assessed is generally representative of the population the program serves. Consider, for example, a nutritional program for families living in poverty designed to improve health and reduce obesity. A comparison group design may include all the families participating in the program at the time of the impact evaluation in the program group, thus ensuring some measure of external validity. The evaluator then faces the challenge of recruiting or constructing a comparison group that will allow internally valid program effect estimates to be derived.

Internal validity for this evaluation would be more readily ensured with a design that controlled the assignment of eligible program participants to treatment and control conditions, but in most circumstances doing so would

be unethical without the permission of the individuals involved. For the nutrition program, eligible families willing to volunteer for a procedure that may sort them into a control group that does not participate in the program will quite likely be different in important ways from typical program participants. For example, they may be less needy and less concerned about nutrition, and thus less bothered by the prospect of being assigned to the control group. Although the internal validity of the resulting program effect estimates might be high for the participants who willingly volunteer for random assignment, their differences from the typical program participants make external validity questionable.

Comparison group designs will not always have external validity advantages relative to the more controlled designs discussed in the next chapter, but it is a factor that an evaluator should consider when designing an impact evaluation. Perhaps the greatest advantage of comparison group designs, however, is that they are generally easier to implement and more convenient than better controlled designs, largely because they do not involve the procedures required to produce that higher level of control and the associated internal validity benefits. Using comparison group designs in an impact evaluation thus requires something of a balancing act. Their relative ease of implementation and potential external validity advantages must be weighed against their greater vulnerability to bias that can compromise internal validity and make the resulting conclusions about program effects misleading if not simply wrong.

With this balancing act in mind, we turn to a discussion of the four types of comparison group designs that are workhorses of impact evaluation practice today and doubtless will continue to be in the future. The major challenge presented by these designs is implementing them in ways that minimize the potential for bias in the effect estimates they generate so that the results provide reasonably credible conclusions about program impacts, and that is the emphasis in the remainder of this chapter.

Comparison Group Designs for Impact Evaluation

Evaluators have used different terms to describe the impact evaluation designs we have referred to here as comparison group designs. More generally, designs such as these are often referred to as **quasi-experiments**, and sometimes as observational studies or nonrandomized designs. It is important to recognize that these types of designs have been under development for more than 50 years and are still being refined and tested today. The term *quasi-experiment* was coined in a classic book for program evaluators titled *Experimental and Quasi-Experimental Designs for Research* by Donald Campbell and Julian Stanley (1963), who wrote,

There are many natural settings in which the research person can introduce something like experimental design into his scheduling of data collection procedures (e.g., the *when* and *to whom* of measurement), even though he lacks the full control over the scheduling of the experimental stimuli (e.g., the *when* and *to whom* of exposure and the ability to randomize exposures) which makes a true experiment possible. Collectively, such situations can be regarded as quasi-experimental designs. (p. 34)

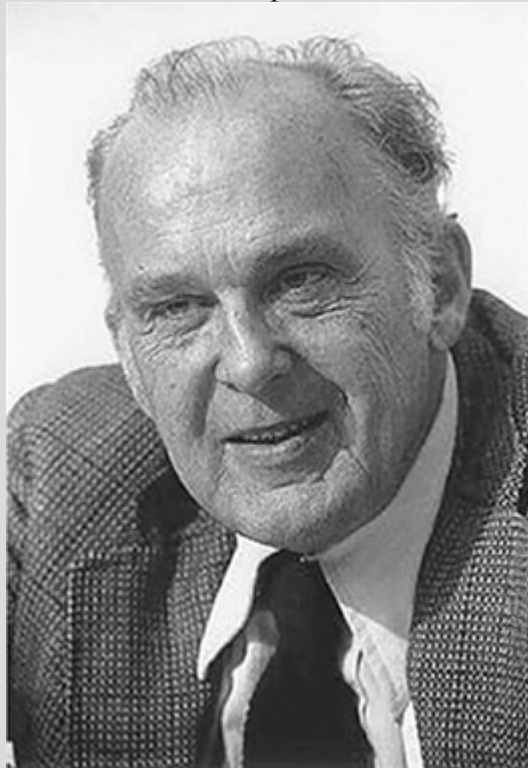
In successive versions of this volume (Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002), as well as in other works, Don Campbell and his colleagues made indelible contributions to program evaluation and social science research generally (see [Exhibit 7-B](#) for a brief biography of Don Campbell). The common characteristics of quasi-experimental designs, as indicated in the quotation above, is that evaluators can control who, when, and what is measured but not exposure to the program. In contrast, the distinguishing characteristic of randomized experiments, which are described in [Chapter 8](#), is that they do exercise control over program exposure as well as who, when, and what is observed in the study.

The comparison group designs reviewed below include naive effect estimates, covariate-adjusted regression effect estimates, matched comparisons, and interrupted time series, with the discussion of the latter category including cohort designs, difference-in-differences, comparative interrupted time series, and fixed effects. As emphasized from the beginning in Don Campbell's writings, quasi-experimental designs of this sort require close attention to the sources of possible bias if they are to produce valid estimates of program effects.

Naive Program Effect Estimates

A naive effect estimate is what results when the average outcome for a group that participated in or had access to a program is simply compared with the average for another group that did not participate in the program or have access to it. The outcome measures may come from administrative data, such as student test scores or length of stay for hospital patients, from surveys that include both treated and untreated individuals, or from direct assessments conducted by the evaluators. But naive effect estimates involve no consideration of the potential for selection bias or other sources of bias that may influence the estimates.

Exhibit 7-B Don Campbell: Evaluation Pioneer and Methodologist



Source: <http://jsaw.lib.lehigh.edu/campbell/obituary.htm>

A notable quotation from Campbell and Stanley's (1963) *Experimental and Quasi-Experimental Designs for Research*:

Internal validity is the basic minimum without which any experiment is uninterpretable: Did in fact the experimental treatments make a difference in this specific experimental instance? External validity asks the question of generalizability: To what populations, settings, treatment variables, and measurement variables can this effect be generalized? Both types of criteria are obviously important, even though they are frequently at odds in that features increasing one may jeopardize the other. While internal validity is the *sine qua non*, and while the question of external validity, like the question of inductive inference, is never completely answerable, the selection of designs strong in both types of validity is obviously our ideal. (p. 5)

A faculty member over the years at The Ohio State University, the University of Chicago, Northwestern University, and Lehigh University, Don Campbell's field of study was scientific inquiry itself, but he was also interested in the use of evaluation research for improving social conditions. Campbell made many contributions in his 40-year career, coining terms such as *quasi-experiment*, *internal validity*, and *external validity*. His books with Julian Stanley, Thomas D. Cook, and William Shadish were considered the field guides for generations of researchers conducting impact evaluations. These books focused evaluators, and social scientists more generally, on the threats to the validity of the effect estimates from research designs for causal inference and provided thoughtful ways to assess those threats. Campbell's methodological contributions flowed largely from his exploration of the philosophy and sociology of science that culminated in his work on "evolutionary epistemology," a distinctive framework for understanding the nature and development of knowledge.

Sources: Brewer (1996) and Thomas (1996).

We share an example of a naive effect estimate here not to guide practice but to illustrate the issues with such estimates. In a study of the effects of universal prekindergarten (pre-K) in Georgia, evaluators measured the outcomes of four groups of students at the end of first grade: (a) former public pre-K attendees, (b) former Head Start attendees, (c) former private preschool attendees, and (d) first graders who did not attend any preschool program (Henry et al., 2005). An evaluator might consider estimating the effect of public pre-K by comparing outcomes for the former pre-K attendees with those of the children who did not attend any preschool. On a measure of math skill, the pre-K children scored 26.3, while the children who did not attend preschool scored 27.0. The naive program effect estimate is the -0.7 -point difference that indicates a small negative effect from attending a public pre-K program. Similar naive estimates of the effects of public pre-K in comparison with private preschool and Head Start yielded score differences of -1.2 and 2.3 points, respectively.

However, the study also found that these four groups of children came from households that were quite different. For example, 29% of the mothers of the pre-K attendees had college degrees, compared with 40% of the mothers of children with no preschool, 48% of the mothers of private preschool attendees, and 4% of the mothers of the Head Start attendees. Moreover, mother's level of education was closely associated with student's score on the math outcome measure. This is a clear case of selection bias: initial differences between the groups on a characteristic other than program exposure that affects the outcome. The differences on the math outcome measure come at least in part from whatever differences in children's math skills are associated with having mothers of different educational levels rather than entirely from the effects of the different preschool experiences. And mothers' education is only one of any number of variables that might create selection bias in this example. As a result, the naive effect estimates cannot be taken at face value as indications of the differential effects of these various preschool experiences; they are biased and quite misleading.

Our purpose in presenting this example of naive estimation is to sound a cautionary note. In many circumstances it may seem to evaluation sponsors that a naive estimate of this sort would provide a low-cost impact evaluation. After all, it compares outcomes for those who received the program with outcomes for those who did not, a comparison that seems as if it should reveal the effects of the program. But it is not an apples-to-apples comparison capable of providing unbiased program effect estimates unless the two groups are so comparable in all the ways that they can be expected to have the same average outcomes in the absence of any program effect. When there are potentially influential differences between the groups, such as having mothers of varying educational levels, it is naive to attribute the simple differences in outcomes to the effects of the program alone.

Covariate-Adjusted, Regression-Based Estimates of Program Effects

In one of the most common comparison group designs used by evaluators, outcomes for a group exposed to a program are compared with those for a comparison group selected on the basis of relevance and convenience. But in contrast to the naive design, this design uses statistical techniques to adjust for differences between the groups that might bias the effect estimates. The first step in this approach is to measure a set of preintervention baseline characteristics for all the members of the study sample, focusing especially on characteristics expected to be related to the outcomes of interest. In this context, these variables are generally referred to as **covariates**. Those covariates are then used in a statistical prediction model that estimates the independent relationship of each covariate to a target outcome variable, that is, what each covariate contributes to the prediction of the outcome above and beyond the contributions of the other covariates included in the statistical model. The statistical model generally used for this purpose is multivariate regression, a well-known form of statistical analysis widely available in standard statistical software packages.

The logic of the covariate-adjusted, regression-based approach to estimating program effects is based on the assumption that whatever part of the differences in the outcome scores is predictable from baseline covariates cannot be a program effect. The difference between the program and comparison group outcomes that remains after adjusting for the influence of the covariates is then assumed to be a less biased estimate of the actual program effect. Of course, any influential covariate omitted from this analysis can still bias the adjusted effect estimate.

The logic of covariate adjustment can be illustrated with a simple example. [Exhibit 7-C](#) presents the outcomes of a hypothetical impact assessment of a vocational training program for unemployed men between the ages of 35 and 40 that was designed to upgrade their skills and enable them to obtain higher paying jobs. A sample of 1,000 participants was interviewed before

they entered the program and again 1 year after it ended. Another 1,000 men in the same age group who did not participate in the program were sampled from the same metropolitan area and also interviewed at the time the program started and 1 year after it ended. In Panel I of [Exhibit 7-C](#), the average post-training wage rates of the two groups are compared without application of any statistical adjustments. Program participants were earning an average of \$7.75 per hour compared with \$8.20 for those who had not participated—not an encouraging contrast for the program. To the extent that participants and nonparticipants differed on characteristics related to earnings other than participation in the program, however, these unadjusted comparisons include selection bias and could be misleading.

Panel II takes one such difference into account by presenting average wage rates separately for men who had not completed high school and those who had. Note that 70% of the program participants had not completed high school compared with 40% of the nonparticipants. When we adjust for the difference in education by comparing the wage rates of persons of comparable educational attainment, the hourly wages of participants and nonparticipants approach each other: \$7.60 and \$7.75, respectively, for those who had not completed high school, and \$8.10 and \$8.50 for those who had. Correcting for the selection bias associated with the education difference thus diminishes the differences between the wages of participants and nonparticipants and yields better estimates of the program effect.

Panel III takes still another difference between the intervention and comparison groups into account. Because all the program participants were unemployed at the time of enrollment in the training program, it is most appropriate to compare their outcomes with those of nonparticipants who were also unemployed when the program started. In Panel III, nonparticipants are divided into those who were unemployed and those who were not at the start of the program. This comparison shows that program participants subsequently earned more at each educational level than comparable, initially unemployed nonparticipants: \$7.60 versus \$7.50, respectively, for those who had not completed high school, and \$8.10 versus \$8.00 for those who had. Thus, when we statistically adjust for the selection bias associated with differences between the groups on education and unemployment, the vocational training program shows a positive

program effect, amounting to a \$0.10/hour increment in the wage rates of those who participated.

Exhibit 7-C Simple Statistical Adjustments in an Evaluation of the Impact of a Hypothetical Employment Training Program

I. Outcome comparison between men ages 35 to 40 who completed the training program and a sample of men ages 35 to 40 who did not attend the program

	Participants	Nonparticipants
Average wage rate	\$7.75	\$8.20
n	1,000	1,000

II. Comparison after adjusting for educational attainment

	Participants		Nonparticipants	
	Less Than High School	High School	Less Than High School	High School
Average wage rate	\$7.60	\$8.10	\$7.75	\$8.50
n	700	300	400	600

III. Comparison after adjusting for educational attainment and unemployment at the time the program began

	Participants		Nonparticipants			
	Less Than High School	High School	Less Than High School		High School	
	Unemployed	Unemployed	Unemployed	Employed	Unemployed	Employed
Average wage rate	\$7.60	\$8.10	\$7.50	\$7.83	\$8.00	\$8.60
n	700	300	100	300	100	500

In any actual evaluation, additional covariates that may differ between the groups and relate to differences in the outcomes would be entered into the analysis. In this example, previous employment experience and wages, marital status, number of dependents, and race might be added—all factors known to be related to wage rates. Even so, we would have no assurance that adjusting for the influence of all these covariates would completely remove selection bias from the estimates of program effects, because influential but unadjusted differences between the intervention and comparison groups might still remain.

Multivariate Regression Techniques

The adjustments shown in [Exhibit 7-C](#) were accomplished in a simple way to illustrate the logic of statistical controls. In actual application, the evaluator would generally use multivariate regression models to adjust for a number of covariates simultaneously. Although multivariate regression is not the only technique that can be used for this purpose, it is by far the most common. The covariates that are important to include in these statistical control models are generally of two different types (see Morgan & Winship, 2014, for the theory underlying this approach). One type has to do with differences between program and comparison groups on preintervention characteristics related to the outcome of interest. For instance, educational level is such a variable in [Exhibit 7-C](#). Other things equal, participants with more education at the beginning of the study are expected to have higher wages at the end. The most important covariates of this sort are preintervention baseline measures of the outcomes that will then be used after the intervention to assess program effects. Preintervention outcome measures are generally the best predictors of postintervention outcomes and thus can be very effective covariates for adjusting for the influence of initial differences on program effect estimates.

The second type of covariate evaluators should seek to identify and incorporate in the analysis relates to differences between the program and comparison groups in terms of their reaction to the program. These covariates adjust for characteristics associated with selection into the program and responses to the program experience for which there may be initial differences between the program and comparison groups. Covariates of this type can be difficult to anticipate in advance and may elude evaluators during the planning process. The influence of motivation on outcomes illustrated earlier in [Exhibit 7-A](#) is an example of such a covariate. Other examples might include such factors as how close individuals live to the program site, how interested they are in participating in the program, or whether they have the characteristics program personnel use to select eligible participants. The importance of variables such as these lies in the fact that if we could fully account for the characteristics that caused an individual to be selected for one group versus the other *and* that also affect the program outcomes, we could statistically adjust for those characteristics and offset the selection bias.

Three additional points are important for identifying covariates with potential to reduce selection bias in any particular comparison group study. First, note that the evaluator's goal is not to completely explain all the variation in the outcomes for the units in the evaluation, or to completely explain selection into the program. The more limited goal is to identify, measure, and model the variation in the outcomes related to differences between the program and comparison groups. For example, the individuals in the program and comparison groups may be completely equivalent with regard to meeting the eligibility requirements for the program, or for how close they live to the program site. Although these characteristics are relevant to the likelihood of participating in the program, and may be related to outcomes, they are not necessary covariates for adjusting bias, because they do not differ for the program and comparison groups and thus cannot create bias.

Second, it is useful in selecting covariates to recognize that each covariate included in the regression model adjusts not only for differences associated with that specific covariate, but also for any other covariates that are substantially correlated with it. Thus, a relevant covariate that is omitted from the analysis will not be problematic if it is highly correlated with a covariate that is included in the analysis. The influence of any covariate is limited to what it can contribute that is not redundant with all the other covariates in the statistical model. For comparison group designs, this means that it is important to prioritize inclusion of strong covariates related to outcomes, such as baseline measures of those outcomes, and strong covariates related to differential selection into the program and comparison groups. With such covariates included, problematic omitted variables will be limited to those that differentiate the groups, are related to outcomes, *and* have no or modest correlations with the covariates already included. Under favorable circumstances, that could be a small set.

The third consideration involves the neutrality of covariates that are redundant with those already in the regression model in a different way. A comparison group that is geographically, culturally, and demographically similar to the program group will already be balanced on many unmeasured characteristics associated with those broad similarities as well as on those factors themselves. Selecting a comparison group with this kind of broad

groupwise similarity to the program group, therefore, will also help reduce selection bias. An impact evaluation of a program in a community mental health center using a comparison group design, for example, would be wise to select a comparison group from the most culturally and demographically similar individuals in the closest geographical proximity possible. Selecting the comparison sample in this way in combination with careful identification and collection of covariates related to selection into the program and the outcomes can reduce bias in the program effect estimates.

Program Effect Estimates From Matched Comparison Groups

Another procedure for reducing bias in comparison group designs is **matching**. In a matched comparison, the intervention group is typically specified first, and the evaluator then constructs a comparison group by selecting units unexposed to the intervention that match those in the intervention group on selected characteristics. The logic of this design requires that, to eliminate selection bias, the groups must be matched on any characteristics that would cause them to differ on the outcome of interest under conditions when neither received the intervention. To the extent that the matching fails to equate the groups on some characteristic that will influence the outcome beyond those on which the groups are already matched, selection bias will remain in the resulting program effect estimate.

Choosing Variables to Match

The first challenge for an evaluator using a matched design is identifying the characteristics that are essential to match. The evaluator should make this determination on the basis of prior knowledge of characteristics related to the outcomes of interest and an understanding of the circumstances that have sorted individuals into program and comparison groups. Relevant information will often be available from the research literature in substantive areas related to the program. For a program designed to reduce pregnancy among unmarried adolescents, for instance, research on teens' risky sexual behavior could be consulted to identify motivations for engaging in sexual behavior, factors leading to early pregnancy, and so on. The objective in constructing a matched comparison group would be to select youth who match those the program is designed to focus on as closely as possible on the important correlates of teen pregnancy.

Special attention should also be paid to identifying variables potentially related to the selection processes that divide individuals into program participants and nonparticipants. For example, in an evaluation of a job

training program for unemployed youth, it might be important to match on their attitudes toward training and their belief in its value for obtaining employment. Even when the groups cannot be matched on variables related to selection, the evaluator should still identify and measure those variables. This allows them to be incorporated into the data analysis to explore and, perhaps, statistically adjust for any associated selection bias by combining matching with covariate-adjusted regression as described in the previous section.

Fortunately, it is not usually necessary to match the groups on every factor mentioned in the relevant research literature that may relate to the outcomes of interest. As described earlier with regard to covariate selection for statistical adjustments, the pertinent characteristics will often be correlated and, therefore, somewhat redundant. For example, if an evaluator of an educational intervention matches students on intelligence measures, the individuals will also be fairly well matched on grade point averages, because intelligence test scores and grades are rather strongly related. The evaluator should be aware of the correlations between the potential matching variables, however, and attempt to match on all the influential factors that are not redundant. If the groups end up differing much on any characteristic that influences the outcome, the result will be a biased estimate of the program effect.

Exact Matching and Propensity Score Matching

Matched comparison groups may be constructed through either exact matching on the selected covariates or matching on propensity scores created from those covariates. In exact matching, the objective is to select a “clone” for each member of the program group from the pool of comparison group members. For children in a school drug prevention program, for instance, the evaluator might want to match on age, sex, number of siblings, and father’s occupation. The evaluator would then scrutinize the roster of children at, say, a nearby school without the program to identify a child with the same profile on these characteristics to match with each child in the drug prevention program. In such a procedure, the degree of closeness may be adjusted to make matching possible—for example, matching within 6 months of age rather than the same month.

Other variants involve departures from one-to-one matching, for instance, matching multiple comparison individuals to each program participant or vice versa. [Exhibit 7-D](#) provides an example of exact matching in a major international study.

The limitations of exact matching as a comparison group design stem mainly from the difficulty of finding exact matches to each program participant on all the covariates the evaluator would like to match on. Potential matches generally should meet any eligibility requirements for program participation, and those individuals must then have data available for all of the covariates that will be used for matching. In many cases a sample of such individuals with the requisite data is not readily available. When such a sample is available, it still may be difficult to find exact matches on the full profile of covariates when matches on a relatively large number of covariates are needed.

An alternative to exact matching that has substantial advantages, and is now widely used in comparison group designs, is **propensity score** matching (Stuart, 2010). In this approach, the program participants and the individuals selected as potential matches are first combined in a common data set, and all the covariates of interest are used in a variant of a regression model that attempts to predict who is a program participant. Usually this is done with a technique known as logistic regression that is especially appropriate for predicting binary outcomes such as participant versus nonparticipant. That analysis yields an estimate of the probability that each individual is in the program group on the basis of the covariates that best differentiate the two groups, that is, the propensity to be in the program group. Those probabilities, ranging from 0 to 1 for each individual in the combined sample, are the propensity scores that can then be used in a matching procedure. Matching on propensity scores produces untreated matches that have similar probabilities of being exposed to the program as the treatment group on the basis of the variables used to predict the propensity of treatment.

Exhibit 7-D Estimating the Effects of a Contingent Cash Benefit Program in India Using a Matched Comparison Group

In 2005, India accounted for 31% of the world's neonatal deaths and 20% of its maternal deaths. To combat these extraordinary death rates, the Bill and Melinda Gates Foundation funded a contingent cash benefit program that paid expectant mothers if they delivered their babies in an accredited medical facility and paid community health workers if they assisted expectant mothers to deliver in such a facility.

Using data from a public health survey, the evaluators identified women of childbearing age who had given birth just after the cash benefit program began. The women who reported receiving the cash benefit were then matched with women who did not report that benefit on state of residence, urban or rural location, below-poverty-line status, wealth, caste, education, number of prior childbirths, and maternal age. With additional covariates used for statistical adjustment (e.g., household distance from the nearest health facility), the evaluators estimated the difference between the program participants and the matched sample that did not participate on the target outcomes.

The results showed that program participants were 43.5% more likely than the matched nonparticipants to have delivered their babies in a health facility, with neonatal deaths reduced by 2.3 per 1,000 live births. The evaluators used two other approaches to create a comparison sample that produced similar effect estimates. Nonetheless, aware of the limitations of comparison group designs, the authors noted that their estimates of the program effects were "limited by unobserved confounding and selective uptake of the programme in the matching [analysis]" (p. 2021).

Source: Adapted Lim et al. (2010).

Propensity score matching typically begins by comparing the distributions of the propensity scores for the program and comparison groups. When there are regions at the tails of those distributions where there are no matches, individuals in those regions may be removed from the analysis. This happens when no members of the comparison group show as high a propensity to be in the program group as some actual members of the program group. Similarly, at the other end of the distributions, there may be no members of the program group who show as low a propensity to be in the program group as some members of the comparison group. Another check on the effectiveness of the propensity matching is to compare the propensity-matched groups on each of the key covariates to ensure that they are equivalent, a condition referred to as covariate balance.

Once created and trimmed and balanced as needed, there are three common ways to actually use the propensity scores to estimate program effects: stratification, weighting, and regression. Stratification is one of the most often used approaches. It typically involves dividing the propensity score distribution into a number of intervals, such as deciles (10 groups of equal overall size), with members of the participant and comparison groups

within each decile, therefore, necessarily having about the same propensity score. Estimates of program effects can then be made separately for each decile group and averaged into an overall effect estimate.

Weighting with propensity scores is done by assigning each member of the participant sample the estimated probability of participating in the program as a weight, and assigning each member of the comparison sample one minus their probability of being in the program as a weight. Weighted averages on an outcome variable are then computed for the program and comparison groups, and the difference in those weighted averages is the estimated program effect. A third way in which propensity scores can be used to reduce selection bias is to simply include them as a covariate in a regression model of the sort described in the earlier section on covariate-adjusted, regression-based program effect estimates. With that method, other individual covariates may also be included in the model (e.g., any of those used to create the propensity scores that were not fully balanced in the results). There are also ways to include individual covariates when propensity scores are used for stratification or weighting, which may further improve the ability of the analysis to reduce selection bias. [Exhibit 7-E](#) presents an example of the use of propensity scores in a comparison group design.

Exhibit 7-E(A) Do Speed Cameras Reduce Traffic Accidents?

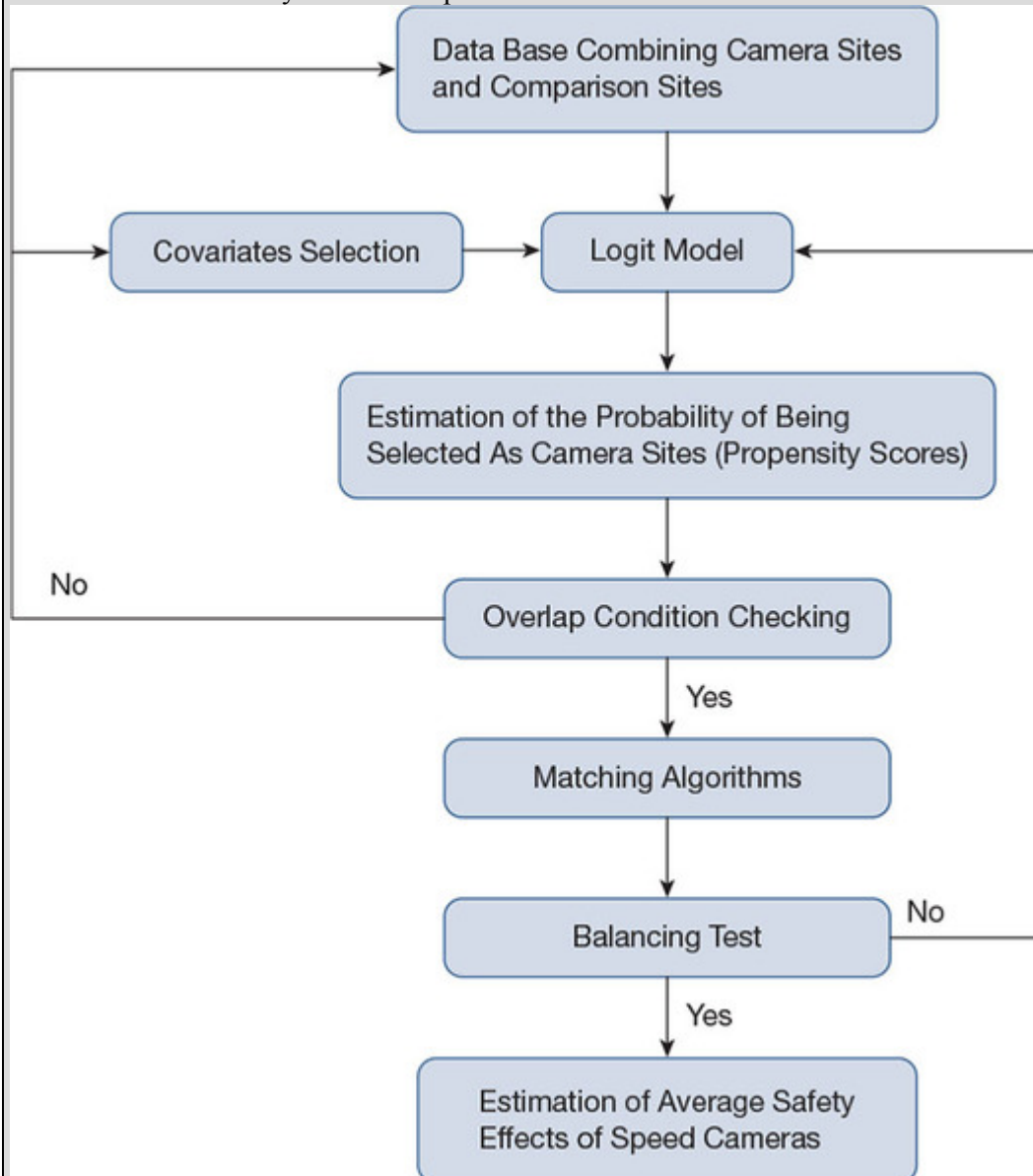
Studies have shown that the number of traffic accidents declines after the installation of cameras that record the license plates of speeding cars for ticketing. However, most of these studies compare traffic accidents after the cameras were installed with those immediately before installation. That comparison is vulnerable to a regression-to-the-mean bias. Speed cameras are often installed in locations where there have been recent increases in traffic accidents, but those increases may be chance outliers after which accident rates would be expected to return naturally to more normal levels for those locations.

Evaluators in England conducted a comparison group study designed to avoid regression-to-the-mean bias. They selected 771 sites where speed cameras had been installed between 2002 and 2004 and used propensity scores to match them on key covariates from a pool of 4,787 potential comparison sites in the same districts without cameras during that period. The evaluators estimated the propensity scores using covariates that included the criteria for selecting a speed camera site and the 3-year traffic accident averages before the installation of any cameras, a period long enough to minimize regression-to-the-mean effects. As shown in [Exhibit 7-E\(B\)](#), they examined the propensity distributions for the program and comparison sites, pruning sites where the scores did not overlap, and

checked the covariate balance to ensure that the propensity score matching was effective in equating the groups on the key covariates.

The results showed that in the range of 500 meters around the sites, fatal or severe accidents were reduced by roughly 16%, and personal injury crashes were reduced by 26%.

Exhibit 7-E(B) Diagram of the Application of Propensity Score Matching to the Evaluation of the Safety Effects of Speed Cameras



Source: Adapted from Li, Graham, and Majumdar (2013).

Propensity score matching has two notable advantages over other matching methods. First, it directly addresses the selection bias issue by focusing on the covariates that show the greatest differences between the program and comparison groups. It is those differences that create selection bias when the covariates are also related to the outcomes of interest, so it makes sense to address selection bias with a method that emphasizes those differences. Second, propensity scores combine information from multiple covariates into a single variable used for matching, often many more covariates than it is practical to use in strategies such as exact matching. At the same time, it is important to recognize that the inclusion of large numbers of covariates may dilute the influence of the most relevant covariates to the point of being counterproductive. In the construction of propensity scores, as with covariates in regression models, addition of covariates highly correlated with another one already included does not improve the performance of the technique for reducing selection bias.

Propensity score matching has become quite popular in recent years. Some of this is due to the flexibility and efficiency of this method for using preintervention covariates to reduce selection bias. But some of the popularity of propensity score methods may reflect a mistaken belief that it is a more complete solution to the problem of selection bias than it may be. It is important to remember that the effectiveness of methods for using covariates to reduce selection bias in comparison group designs is overwhelmingly dependent on including all the relevant covariates. Whether covariates are used in regression models, for direct matching, or in propensity scores, it is always possible that some degree of selection bias remains because of the omission of critical covariates. Although a useful technique, propensity score methods cannot overcome an inadequate set of covariates when an evaluator is trying to remove selection bias in a particular comparison group evaluation.

Interrupted Time Series Designs for Estimating Program Effects

The comparison group designs discussed in this section differ in at least one important way from those reviewed above. Whereas those designs compared the outcomes of two groups—a program group and a comparison group—**interrupted time series** designs compare outcomes for a period before program implementation or participation with those observed afterward. The program or other intervention in these designs “interrupts” a time series of periodic measures of a relevant outcome the program is expected to affect. The threats to the internal validity of these designs are not dominated by the selection bias issue but, rather, relate mainly to factors other than program onset that can bring about change in the series of outcome measures and thus potentially mimic a program effect. Coinciding events, secular trends, maturation, and regression to the mean, for instance, may bias program effect estimates from time series designs.

Because of the need for periodic measures of target outcomes before, during, and after program onset, evaluations using interrupted time series designs most often draw their data from existing databases that track key indicators in such areas as health, crime, education, employment, and the like. In the next four subsections of this chapter, we describe various research designs involving interrupted time series that can be used for impact evaluation. We begin with the cohort design, which is not generally the strongest time series design for minimizing potential bias, but is relatively common and provides the underlying conceptual framing for the more rigorous interrupted time series designs that follow.

Cohort Designs

Cohort designs estimate the program effect by comparing outcomes for the cohort(s) of individuals exposed to a newly initiated or revised program with those for the cohort(s) before that with no such exposure. For example, an organization providing relapse prevention training to smokers who want to quit might add a nicotine patch component to that intervention. The 6-

month relapse rates for some number of cohorts of individuals who went through the program after adding the nicotine patch would then be compared with the 6-month relapse rates for those in some number of cohorts who went through the program before the patch was added to obtain an estimate of the effect of adding the nicotine patch. Or, consider a nurse home visitation program initiated for low-income pregnant women during the prenatal period and the 1st year thereafter. Comparison of infant health indicators for the birth cohorts of children of eligible women before the program was initiated and afterward might then be used to estimate the effects of the program on relevant health outcomes.

Any program that routinely enrolls participants in a time-limited or age-specific service is appropriate for a cohort design assessing the effect of exposure to an intervention if it is possible to obtain outcome measures for successive cohorts before and after that intervention is introduced. The average outcomes in the preintervention period are used to estimate what would have happened in the absence of the intervention (i.e., the counterfactual outcomes). For the resulting program effect estimates to be valid, various sources of potential bias would have to be ruled out or statistically controlled. Interfering events or changes in secular trends around the time the intervention is initiated, for instance, would introduce bias if they affected the outcomes of interest. Similarly, any changes across cohorts that influenced the selection of participants into the program such that they might naturally have different outcomes would introduce bias. In short, any source of change in the observed outcomes other than those associated with program exposure may introduce bias if it occurs close to the time of program exposure, including those related to the outcome data collection or the recordkeeping that is the source of the data.

The inherent vulnerability of cohort designs to such biasing influences is evident in the evaluation of the Massachusetts health insurance reform that became the model for the Affordable Care Act (described in [Exhibit 7-F](#)). The evaluators did not rely on a cohort design for their effect estimates, but their report provides the information they would have used if the cohort design had been implemented. The evaluators examined data on self-reported physical and mental health during at least 28 days of the previous month for adults in Massachusetts before and after the insurance reforms

were introduced. They used covariate adjustment techniques like those described earlier to account for differences associated with variables such as sex, age, income, and the state unemployment rate. The covariate-adjusted percentage reporting good physical health showed a slight increase from 79.8% to 80.4%, with the good mental health percentages showing an even smaller increase from 75.1% to 75.2%. However, the evaluators recognized a potential source of bias in these estimates related to a change in the methods of data collection after the reform began that would have compromised a simple cohort design. Instead, they used a more sophisticated difference-in-differences design ([Exhibit 7-F](#)) that provided a more credible effect estimate, which we describe in the next section.

Difference-in-Differences Designs

Difference-in-differences designs are interrupted time series designs that compare pre- and postintervention outcomes in sites that implemented the intervention to analogous before-after changes in sites in which it was not implemented, thus adding a comparison time series to the intervention one. For present purposes, we view this design as involving outcomes in the period immediately preceding the introduction of the intervention and those in the immediately following period. With longer pre- and postintervention time periods, trends in the respective outcomes for intervention and comparison time series can be examined. Those designs, referred to as comparative interrupted time series designs, are discussed in the next section.

Exhibit 7-F Evaluating the Effects of the Massachusetts Health Care Reform of 2006: An Example of a Difference-in-Differences Design

In 2006, Massachusetts sought to improve the health of its residents by expanding health insurance coverage. The state required that residents obtain health insurance, expanded Medicaid coverage, subsidized health insurance for lower income residents, and established a health insurance exchange to facilitate access to insurance. Implementation was successful, as evidenced by the fact that immediately after this reform Massachusetts had the highest rate of health insurance coverage (98%) and the greatest gains in coverage in the United States for low-income residents.

The evaluators used public health survey data collected in Massachusetts and five other New England states with no insurance changes to estimate the effects of the Massachusetts reform. Data from 2001 through 2006 provided health-related outcomes

for the prereform period and data from 2007 through 2011 provided the postreform outcomes. The difference-in-differences design used by the evaluators examined the extent to which the before-after differences in Massachusetts exceeded the before-after differences in the other New England states that provided the comparison time series. [Table 7-F1](#) shows the difference-in-differences effect estimates for the outcomes examined.

The importance of the comparison group of other New England states is evident in the first row of [Table 7-F1](#). Massachusetts residents reporting excellent or very good health declined from the prereform to postreform period by 0.7 percentage points, but the decline of 2.4 percentage points in the comparison states was even larger. The difference in these differences thus showed a 1.7 percentage point advantage for Massachusetts. Overall, the health of residents of Massachusetts improved relative to that of the residents of the comparison states for 9 of the 10 health-related outcomes.

Table 7-F1

TABLE 7-F1 Adjusted Probabilities Before and After Massachusetts Health Reform

	Massachusetts		Other New England States		Difference-in-Differences
	2001–2006	2007–2011	2001–2006	2007–2011	
Health status					
General health	66.2	65.5	66.6	64.2	1.7*
Physical health	79.8	80.4	80.1	79.4	1.3*
Mental health	75.1	75.2	75.6	74.2	1.5*
Preventive health care access					
Mammogram	85.3	85.6	83.2	83.6	−0.1
Pap test	93.8	93.3	93.5	90.7	2.3*
Colonoscopy	59.7	71.2	61.8	67.8	5.5*
Cholesterol check	90.7	92.6	90.2	90.7	1.4*
Health care access					
Covered by insurance	94.4	96.8	92.9	91.7	3.6*
Have personal physician	89.7	90.7	88.4	87.8	2.4*
Have cost barriers	5.9	5.8	7.5	9.4	−2.0*

*Statistically significant differences.

Source: Adapted from Van der Wees, Zaslavsky, and Ayanian (2013).

Statistically significant differences.

Source: Adapted from Van der Wees, Zaslavsky, and Ayanian (2013).

The advantage of the difference-in-differences design relative to the simple cohort design is the inclusion of before-and-after outcomes for comparison sites where there was no exposure to the intervention. Before-after differences in those sites, of course, cannot be intervention effects and must, therefore, represent some other source of change—one that might bias the before-after difference in the intervention sites. By essentially subtracting out that presumptively biasing difference in the comparison sites from the

difference in the intervention sites, we get a difference between the differences that should be a less biased estimate of the intervention effect. [Exhibit 7-F](#) describes a difference-in-differences design that assessed health-related outcomes for Massachusetts residents before and after legislation that increased health insurance coverage. The comparison sites were other New England states that made no changes in health insurance and were thus used to represent what would have occurred in Massachusetts in the absence of the reform legislation.

To consider program effect estimates from a difference-in-differences design as plausibly unbiased, several assumptions must hold. These can be illustrated by reference to the Massachusetts evaluation summarized in [Exhibit 7-F](#). First, conditional on the covariates used to adjust the estimated differences, the basis for selection into the study samples must be the same before and after the time when the intervention is introduced. The evaluators in the Massachusetts study used covariate-adjustment techniques such as those described earlier to equate the before and after samples on demographic characteristics such as age and education as well as state-level unemployment rates. Other secular trends that might have produced before-after changes in the health of Massachusetts residents were further controlled by subtracting out the before-after changes observed in the comparison states. Although possible, it seems somewhat unlikely that there were changes in the health of Massachusetts residents aside from the effects of the reform that would not have also occurred in neighboring states with their similar populations and health care trends. And, indeed, graphs of the preintervention trends in Massachusetts and the comparison states presented by the researchers demonstrated that they were comparable.

Another threat to the internal validity of any time series design is the concurrence of other events with the initiation of the intervention that might influence the target outcomes. When a major legislative change such as the insurance reform in Massachusetts is made, it is not unusual for other initiatives to also be launched or under way that relate to the same concerns that motivated the intervention being evaluated. There was no report of any such coinciding events that would plausibly affect the health of Massachusetts residents for the example used here. To be confident that it is the focal intervention that has caused the observed effects, evaluators using

any time series design must have sufficient awareness of other concurrent events to conclude that none were plausible alternative explanations for any changes observed.

Another consideration, as mentioned for cohort designs, is regression-to-the-mean bias that enters the time series when the intervention being tested is implemented after an atypical adverse spike in the target outcome. In the Massachusetts insurance example, the evaluators averaged over multiple years of data from the preintervention period to reduce the likelihood that the before-after change observed was the result of outlier values on the health indicators immediately before the insurance reform. In addition, it is wise for evaluators to examine the preintervention trend in the outcome measures in both the reform and comparison groups to identify any atypical values that might signal the potential for a regression-to-the-mean bias.

Comparative Interrupted Time Series Designs

Comparative interrupted time series designs are similar in terms of their underlying logic to difference-in-differences designs, but they include sufficient preintervention data to model the trend over time that leads up to the onset of the intervention. This allows the intervention effect to be estimated as a deviation from that preintervention trend rather than relying on a more compressed before-after comparison. To implement a comparative interrupted time series design, at least four periods of data are needed before the intervention, and more may be necessary if the trend does not take a simple form or there is great variability in the data points around the underlying trajectory. The comparative aspect of this design, like difference-in-differences designs, involves a time series in a similar context in which there is no exposure to the intervention. That time series then also provides a preintervention trend line that, ideally, should be comparable with that for the intervention series, and that allows before-after trends without intervention to be incorporated into the estimate of the effect with intervention.

In one example of a comparative interrupted time series design, an evaluation of the effects of the federally funded Reading First program was conducted for the participating schools within one state (Jacob, Somers,

Zhu, & Bloom, 2016). Reading First provides kindergarten through third grade support for reading curricula and materials that meet federal standards and associated professional development and coaching for teachers. The evaluators obtained school-level reading test scores and other data from publicly available databases for the 6 years before the intervention and 2 years afterward for elementary schools in the respective state. This 6-year preintervention time series allowed the evaluators to assess the extent to which the postintervention test scores deviated from the preintervention trend. Using schools that did not participate in Reading First as a comparison group, they found no differences in those deviations from prior trends between the Reading First schools and the comparison schools. To explore concerns about the comparability of the comparison schools, the evaluators estimated program effects using three comparisons that are more similar to the treated schools in specific ways: only elementary schools in districts eligible for the program, only schools that applied for the program, and only schools matched on preintervention trends. The results were essentially the same in all these comparisons and the original analysis. In an exceptional further contribution of this particular study, the evaluators were able to compare the results from the comparative interrupted time series with those from a more rigorous design also applied to the Reading First program in the same state (a regression discontinuity design, discussed in the next chapter). The results were substantially similar, lending support to the view that the time series design produced plausibly unbiased program effect estimates in this instance.

Fixed Effects Designs

Fixed effect designs involve time series outcome data for *each* unit within a group of units, at least some of which are exposed to the program at some of the times in the time series and not at others. The average outcome over time for each unit is subtracted from the outcome at each observation period for that unit, and a program effect is estimated as the difference between the deviations from that average for the periods of program exposure and the periods without exposure, adjusted when appropriate for the differences in the time period for the comparison units that were never or always exposed to the program. The overall program effect estimate is then the average of the effect estimates across all the units. The advantage of this design is that

each unit serves as its own control. That is, factors that do not vary for each unit over the course of the time series are necessarily constant and cannot affect the deviations on which the program effect estimates are based. The units in a fixed effects design may be individuals, households, communities, or any other units that may differ in ways that could otherwise bias the effect estimate. Thus the kinds of factors that can be held constant in this design include such things as individuals' innate ability, the education level of adult members of a household unit, urban or rural location of a community, and so forth. Eliminating differences of this sort that occur between units (e.g., innate ability) from influencing the effect estimates can be an effective way to reduce some sources of bias in a comparison group design.

An example will help illustrate the potential value of fixed effects comparisons (Lindo & Packham, 2015). In Colorado in 2008, a private donation funded access to long-acting reversible contraceptives through clinics with federal funding to provide family planning and prevention services for low-income women. Because these contraceptives are expensive, they had not been previously made available by most of the state's clinics, and the use rate by teens was less than 3%. The question for the evaluators was whether increasing access to these contraceptives reduced teen pregnancies.

The basic evaluation design was a comparative interrupted time series that compared before-after changes in the trends for teen pregnancy rates in the Colorado counties served by clinics that received the funding to increase access with those in counties in other states served by comparable clinics supported under the federal program for family planning and prevention services. Data were available for 7 years before the intervention and 4 years after. A complication, however, was the downward secular trend in teenage pregnancy rates across the United States during the period when access in the Colorado counties was expanded. If that downward trend was quite different for the comparison counties than the intervention counties, there was potential for selection bias in the effect estimates based on that comparison. To minimize that potential, the evaluators used a county fixed effects design in which the before-after trend differences were estimated within each county to minimize between county differences on inherent

county characteristics associated with different trends. The results indicated that the initiative to increase access reduced teen birth rates by 4% to 7% over the years after it was implemented.

Of course, as with any design, there are limitations. Because the outcomes for any period are analyzed as deviations from an average, there must be at least two observations per unit so an average can be calculated. In addition, at least some of the units included in the effect estimate must have been exposed to the intervention for one or more observations and not in one or more observations. These units are referred to as switchers, and switchers may not be representative of the target population for the intervention. That may raise questions about the generalizability (external validity) of the effect estimates beyond the subset of units on which they were estimated.

As with any of the interrupted time series designs, fixed effects designs are not inherently capable of eliminating selection bias. However, by adding fixed effects for study units, the between-unit differences that are stable within units, but may be sources of selection bias, are controlled, thus minimizing one source of potential selection bias. More generally, the increased amount of information from preintervention data used in time series designs can improve the estimates of counterfactual outcomes and address such other sources of bias as secular trends and interfering events.

Cautions About Quasi-Experiments for Impact Evaluation

The superior ability of well-controlled, well-executed designs to produce unbiased estimates of program effects, such as the randomized control design described in the next chapter, makes them the obvious choice if they can be implemented within the practical constraints of an impact evaluation. Unfortunately, the environment of social programs is such that those designs can sometimes be difficult or impossible to conduct and implement well. The value of comparison group designs is that, when carefully done, they offer the prospect of providing credible estimates of program effects while being relatively adaptable to program circumstances. Furthermore, some comparison group designs in some circumstances may provide program effect estimates with greater external validity than would be possible within the constraints inherent in a more rigorous design. Better generalizability of a biased estimate of program effects, however, is a dubious advantage, so the ability of comparison group designs to produce effect estimates with acceptable internal validity is still a critical concern.

A central question, therefore, is how good comparison group designs typically are for producing unbiased estimates of program effects. Put another way, how much risk for serious bias does the evaluator run when using quasi-experimental research designs instead of randomized control designs? We would like to be able to answer this question by drawing on a body of research that compares the results of various quasi-experimental designs with those of randomized experiments in different program situations. Such studies are rare, although they are becoming more common. What the available studies that make these comparisons show is what we might expect: Under favorable circumstances and carefully done, comparison group designs can yield estimates of program effects similar to those from randomized designs, but they can also produce quite different and erroneous results.

In an early investigation of this issue, Lipsey and Wilson (1993) compared the mean effect size estimates reported for randomized versus nonrandomized designs within 74 meta-analyses of psychological, educational, and behavioral interventions. In many of the meta-analyses, the

estimates of the effects for the interventions of interest from the nonrandomized designs were similar to those from the randomized designs. However, there were also many instances of substantial differences, with the nonrandomized studies sometimes producing much larger effect estimates than the randomized ones and sometimes producing much smaller ones. Heinsman and Shadish (1996) made a closer examination of the effect estimates in 98 studies within four program areas and also found that nonrandomized designs gave varied results relative to randomized designs—sometimes similar, sometimes appreciably larger or smaller.

More recent investigations, often called validation studies or within-study comparisons, have focused on the conditions under which comparison group designs are most likely to produce program effect estimates similar to those from randomized control designs, keeping as many other factors the same as possible. Shadish, Clark, and Steiner (2008) found that if covariates are available that are correlated with both selection into treatment and the program outcome, then as expected, matching and covariate-adjusted regression both reduce bias. Other similar studies suggest that using a baseline preintervention measure of the outcome as a covariate or for matching generally results in a substantial reduction of bias. Also, selection of the comparison sample from the same geographic area as the program sample may help reduce bias. Finally, ensuring that the comparison group is eligible for the program and, if possible, similarly motivated to participate in the program appears to have benefits for reducing selection bias.

Given all the limitations of comparison group impact evaluation designs pointed out in this chapter, when can their use be justified? Clearly, they should not be used if it is possible to use an inherently more rigorous design. However, when that is not possible and an impact evaluation is needed for good reasons, then conducting the strongest comparison group design feasible for the program circumstances is a reasonable option. It is especially important in that case that the evaluator have an awareness of the limitations of the selected design and make vigorous attempts to overcome them. A responsible evaluator will also advise stakeholders of the limitations of the evaluation design chosen and the confidence that can be placed in the results given those limitations.

Summary

- Impact evaluation aims to determine what changes in outcomes can be attributed to the intervention being evaluated. Although the strongest research designs for this purpose, such as randomized control designs, strictly control access to the program, comparison group designs that do not require control of program access can be used when inherently stronger designs are not feasible.
- A major concern of evaluators in any impact evaluation is the potential for bias that might compromise the validity of the estimates of program effects. Among the possible sources of bias that may be especially problematic in comparison group designs are selection bias, secular trends, interfering events, maturation, and regression to the mean.
- In comparison group designs, outcomes are obtained for individuals or other units that are naturally exposed to the program without any manipulation of their access or opportunity to participate. The distinctive feature of these designs is that the comparison group used to estimate the counterfactual outcomes for the program group is constructed from a pool of individuals who were not exposed, or not yet exposed, to the intervention. This comparison does not ensure that the individuals in the program and comparison group are comparable in the way necessary to support a valid estimate of the program effect. That is, the two groups might not have identical outcomes in the absence of the program or when it has no effect.
- One family of comparison group designs compares outcomes for a group of individuals exposed to the program and a group of different individuals who were not exposed. Preintervention baseline data on selected characteristics of those individuals, referred to as covariates, can be used in various ways to reduce potential bias in the program effect estimate. The covariates most relevant to potential bias are those that show differences between the program and comparison groups and are also related to the outcomes of interest. Bias can remain in the program effect estimate if any such covariate is omitted from the analysis, unless it is largely redundant with those already included.
- One approach to using covariates to reduce bias is to use them in a multivariate regression analysis model that statistically adjusts the effect estimate for influential initial differences between the groups. Another approach is to match individuals in the program group with individuals in the comparison group so that the two groups have the same profile on the selected covariates. An especially efficient and effective way to use covariates is to combine them to create something called a propensity score, which can then be used for matching or in other ways in the analysis to adjust for initial differences on influential covariates.
- Another family of comparison group designs, generally referred to as interrupted time series, compares outcomes for selected units for some period before the introduction of the intervention with those for some period after. Variants of these designs differ mainly on the extent to which they reduce bias from events concurrent with program onset, secular trends, maturation, and regression to the mean.
- Time series designs include simple comparison of outcomes from successive cohorts before and after the initiation of a new or modified program. Difference-in-differences designs add before-after outcomes for comparison units not exposed to the program. Comparative interrupted time series designs also use a comparison timeline, but include repeated measures of the outcomes so that trends, and

discontinuities in those trends associated with the onset of the intervention, can be included in the analysis. Fixed effects designs examine trends and discontinuities in trends for each unit in the sample contributing the time series data, thus eliminating any bias associated with difference between units on characteristics that are stable within a unit.

- Comparison group designs, also known as quasi-experimental designs, often have advantages, including relative ease of implementation and potentially greater generalizability of the program effect estimates (external validity). Because of their greater vulnerability to bias, however, stronger designs should have preference when feasible. When these designs are used, it is essential that the evaluator be aware of the potential for bias, take steps to minimize it as much as possible, and acknowledge the limitation of the design when reporting the results of the evaluation.

Key Concepts

Attrition 160
Comparison group 158
Covariate 168
External validity 164
Interfering event 162
Internal validity 164
Interrupted time series 176
Intervention group 158
Matching 171
Maturation 163
Program effect 158
Program impact 158
Program group 158
Propensity score 172
Quasi-experiment 165
Regression to the mean 163
Secular trends 162
Selection bias 160

Critical Thinking/Discussion Questions

1. Describe the five types of bias discussed in [Chapter 7](#) and provide an example of each type.
2. The first challenge for an evaluator using a matched design is identifying the characteristics that are essential to match. Pick a social intervention to evaluate and identify five variables that are essential to match. Why are these five variables important?
3. Define the four different interrupted time series designs for estimating program effects that are discussed in the chapter. Provide an example of each type of design.

Application Exercises

1. Locate an evaluation report of a large social intervention and determine what kinds of potential bias the researchers had to contend with. What did the researchers do to limit vulnerability to those sources of potential bias? Do you believe those attempts were sufficient for producing unbiased estimates of the program effects?
2. A central question in impact evaluations is how much risk of serious bias the evaluator runs when using quasi-experimental research designs instead of randomized control designs. The discussion in this chapter reports that some research has been conducted that compares the results of various quasi-experimental designs with those of randomized experiments. Locate one of these studies and produce a short summary of its findings.

Chapter 8 Impact Evaluation Designs

With Strict Controls on Program Access

[Controlling Selection Bias by Controlling Access to the Program](#)

[Randomized Control Designs](#)

[Regression Discontinuity Designs](#)

[Key Concepts in Impact Evaluation](#)

[Program Circumstances](#)

[Types of Counterfactuals](#)

[Types of Program Effects](#)

[Unit of Assignment](#)

[Multiple Intervention Conditions](#)

[When Is Random Assignment Ethical and Practical?](#)

[Ethical Considerations](#)

[Practical Considerations](#)

[Application of the Regression Discontinuity Design](#)

[Choosing an Impact Evaluation Design](#)

[Summary](#)

[Key Concepts](#)

Impact evaluations are undertaken to find out whether programs produce the intended effects on their target outcomes. Only evaluations that strictly control access to the program can remove the vulnerability of program effect estimates to selection bias. The two types of impact evaluation designs with these characteristics are described in this chapter: randomized control designs and regression discontinuity designs. Among impact evaluators, it is widely recognized that well-executed randomized designs produce the most methodologically credible estimates of program effects. Evaluations using regression discontinuity designs also have a high degree of inherent internal validity and are generally recognized as second only to randomized designs in terms of the credibility of their program effects estimates.

Although designs that strictly control access to the program are the strongest for eliminating selection bias, implementing them can be challenging and is not always feasible in practice. Also, because of the controls on program access they require, the social benefits expected and

the need for credible evidence about impact must be sufficient to justify the use of these designs.

Choosing a design for an impact evaluation must take into account two competing pressures. On one hand, such evaluations should be undertaken with sufficient rigor to support relatively firm conclusions about program effects. On the other hand, practical considerations and ethical treatment of potential participants in the evaluation limit the design options that can be used.

Although impact evaluations are highly prized for the relevance of their results to deliberations about continuing, improving, expanding, or terminating a program, their value for such purposes depends on the credibility of those results. Impact evaluations that misestimate program effects will make misleading contributions to such discussions. A program effect or impact, as you may recall from previous chapters, refers to a change in the target population or social conditions brought about by the program, that is, a change that would not have occurred without the program. The main difficulty in isolating program effects is establishing a counterfactual: the estimate of the outcome that would have been observed in the absence of the program. As long as a reliable and valid measure of the outcome is available, it is relatively straightforward to determine the outcome for program participants. But it is not so straightforward to estimate the outcome for the counterfactual condition in which those same participants were not exposed to the program. In [Chapter 7](#), we reviewed ways to estimate the counterfactual when not everyone appropriate for a program actually participates, with participation determined more or less naturally by individual choice, policymakers' decisions to make the program available, or administrative or staff discretion. In this chapter, we focus on designs that control access to the program so that the basis for differential program exposure is known in ways that make it possible to avoid the potential for selection bias that plagues the designs described in [Chapter 7](#).

There are two impact evaluation designs that control access to a program in ways that can eliminate selection bias, but they do so in very different ways: **randomized control designs** (also known as randomized control trials,

RCTs, and randomized experiments) and **regression discontinuity designs**. These designs are widely considered the most rigorous options available for impact evaluation.

Controlling Selection Bias by Controlling Access to the Program

All impact evaluations are inherently comparative: Observed outcomes for relevant units that have been exposed to a program are compared with estimated outcomes for the corresponding counterfactual condition. In practice, this is usually accomplished by comparing outcomes for program participants with those of individuals who did not experience the program. Ideally, the individuals who did not experience the program would be identical in all respects except for exposure to the program. The two impact evaluation designs that best approximate this ideal involve establishing control conditions in which some members of the target population are not offered access to the program being evaluated. The *control group* or *control condition* terminology here is used in contrast to the *comparison group* phrasing in [Chapter 7](#) because of the controlled access to the program that creates this group in these more rigorous designs.

Randomized designs and regression discontinuity designs establish **control groups** in ways that differ in their logic and the means through which the control group is created. These designs are not considered completely equal with regard to their vulnerability to selection bias. Well-executed randomized designs are generally recognized as having greater inherent internal validity for the impact estimates they yield. But both these designs offer greater protection against selection bias than virtually all alternative impact evaluation designs. Next we explain the logic and distinct benefits of each of these designs.

Randomized Control Designs

The critical element in estimating program effects by comparing outcomes for an intervention group with those from a control group is configuring the control group so that it is equivalent to the intervention group before any experience with the program. Equivalence, for these purposes, means the following:

Identical composition: Intervention and control groups contain the same mixes of persons or other units in terms of their program-related and outcome-related characteristics.

Identical predispositions: Intervention and control groups are equally disposed toward the program and equally likely, without intervention, to attain any given outcome status.

Identical experiences: Over the period of observation, the intervention and control groups experience the same time-related processes other than the program experience: maturation, secular trends, interfering events, and so forth.

Although perfect equivalence could theoretically be attained by matching each unit in an intervention group with an identical unit that is then included in a control group, this is clearly impossible in program evaluations. No two individuals, families, or other units are identical in all respects. Fortunately, one-to-one equivalence on all characteristics is not necessary. First, it is only necessary for intervention and control groups to be equivalent in aggregate terms; that is, the group averages should be the same. Second, it is only necessary that the groups be equivalent on characteristics, predispositions, and experiences that are related to the program outcomes being evaluated. It may not matter if the intervention and control group members differ in place of birth or favorite color, as long as these differences are not associated with differences on the outcome.

Random assignment, also referred to simply as *randomization*, is the most effective way to ensure the aggregate equivalence of the intervention and control groups in an impact evaluation. Random assignment means that a probabilistic procedure determines whether each individual (or other unit)

in the evaluation sample will be a member of the intervention group or the control group. The result of randomization is that the two groups differ only by chance on virtually everything about them, whether relevant to the outcome (most important) or not, and whether a known concern for the evaluator or not.

Random assignment does not mean that some haphazard, arbitrary, or capricious process was used to assign individuals to groups. On the contrary, random assignment requires that an explicit probabilistic process be used to sort an initial sample of appropriate individuals into the intervention and control groups. Moreover, there must be strict adherence to the results of that procedure so that membership in the respective groups is determined solely by chance. Random assignment, therefore, involves such chance processes as a coin toss, names drawn from a hat, or a roll of dice to determine the group to which each individual in the sample is assigned. As a practical matter, computer-generated random numbers are generally used for this purpose. For example, the sample of eligible individuals might be organized into a list with a column of computer-generated random numbers with a random start laid alongside the list. The random numbers are then used to sort the list, which will then be in random order. If half the individuals are to be assigned to the intervention group, then the first half of that randomly sorted list can be used to identify those individuals, while the second half identifies the control group.

The result of this process is assurance that any difference between the intervention and control groups has occurred literally by chance, not by any systematic sorting of individuals with different characteristics into the groups—the very situation that potentially produces selection bias. Just as chance tends to produce equal numbers of heads and tails when a handful of coins is tossed into the air, chance tends to make intervention and control groups equivalent. Of course, if only a few coins are tossed, the proportions of heads and tails may, by chance, be quite different, the likelihood of which diminishes as the number of coins increases. Similarly, if only a small number of individuals are randomly assigned, problematic differences between the groups could arise, and with bad luck, that might even happen with larger samples—what evaluators call “unhappy randomization.”

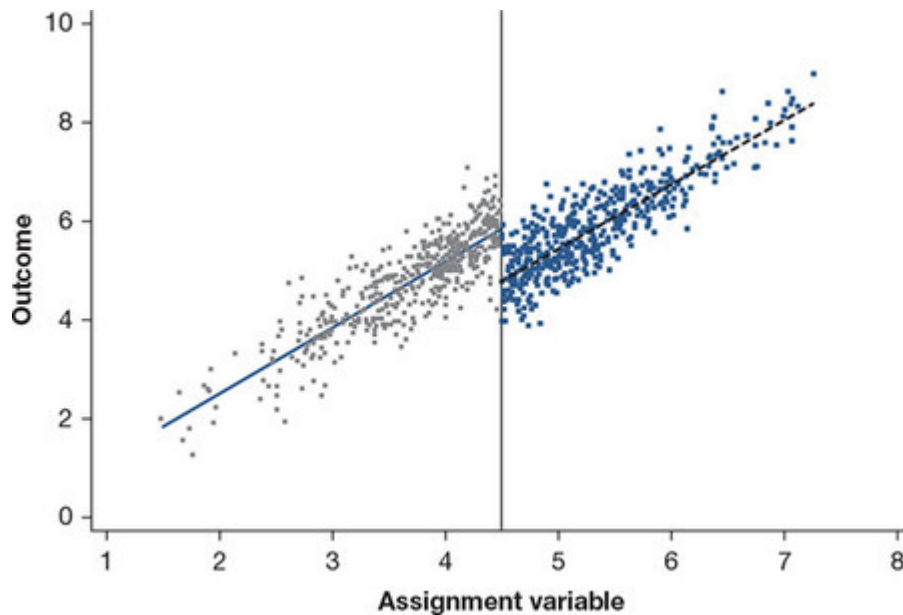
Another advantage stemming from the chance process for random assignment is that the proportion of times that a difference of any given size on any given characteristic can be expected in a series of randomizations can be calculated from statistical probability models. This is the basis for statistical significance testing of the outcome differences between intervention and control groups. Such statistical tests guide a judgment about whether an observed difference on an outcome is likely to have occurred simply by chance or more likely represents a true difference. If the observed difference is expected to occur by chance rather infrequently (less than 5% of the time by convention), the difference in the average outcomes between the intervention and control groups is thus highly likely to represent an intervention effect. [Chapter 9](#) presents a fuller discussion of the statistical framework for impact evaluation designs with varying sample sizes.

Regression Discontinuity Designs

Regression discontinuity designs rely on a **quantitative assignment variable**, also called a forcing variable or cutting-point variable, rather than chance, to assign individuals to the intervention or control group. Like randomized designs, however, the procedure for assigning individuals to groups is part of the research design itself and is thus fully known. Whether chance or the score on an assignment variable controls assignment to treatment or control groups, it is this controlled assignment that accounts for the reduced vulnerability to selection bias of these designs.

For this design, each individual first receives a score on the **assignment variable**, and one score within that range is then designated as the cut point. A strict sorting then assigns everyone scoring below that cut point, even by just a little bit, to one group and everyone scoring above that cut point to the other group. For example, we might measure the reading ability of a sample of third grade students and use that as an assignment variable. A cut point on that measure of reading ability might then be chosen that differentiates the poorest readers who most need assistance from those above that threshold who are less in need of additional reading instruction. The students scoring below the cut point are then assigned to participate in a remedial reading program, and those above the cut point do not participate in that program and serve as the control group. After the remedial reading program is over, outcome reading scores are then measured for both groups. [Figure 8-1](#) shows what a positive effect of the reading program would look like when the scores on the reading outcome are plotted against the scores on the reading assignment variable.

Figure 8-1 A Cut Point (4.5) on the Variable That Assigns Units to the Treatment or Control Group, With Those Below the Cut Point Receiving an Intervention That Boosted Their Scores on the Outcome Measure



Gray denotes the treatment group, and blue denotes the control group.

The critical area in a regression discontinuity plot like [Figure 8-1](#) is the interval on the assignment variable that is right around the cut point. Individuals just barely above that cut point and those just barely below have been differentiated only by small differences in their scores on the assignment variable. As such, they can be expected to be similar in all respects except that those on one side have access to the program while those on the other side do not. For this to be true, the cut point has to be set on the basis of criteria that are unrelated to the outcomes. For example, the assignment variable might be a measure of risk for some adverse outcome collected at baseline, with the cut point for assignment to a prevention program set according to the number of individuals the program can serve. Or the assignment variable might be a measure of need, with the cut point determined by the eligibility criteria for a program that serves clients judged to most need their services.

If the location of the cut point is determined on the basis of such independent considerations, the individuals close to the cut point on one side and those close to the cut point on the other side are effectively randomized except for any influence on the outcome of their small differences on the assignment variable. However, any relationship of the

assignment variable to the outcome can be statistically controlled, for example, by treating it as a baseline covariate in a regression analysis, as described in [Chapter 7](#). Because the assignment variable is known to be the sole basis for selection into intervention and control groups, there are no other potential sources of selection bias once it is controlled. With that done, any difference on the outcome can be attributed to the program; that is, it is an estimate of the program effect.

The evaluator must determine how far from the cut point it is reasonable to go with confidence that the outcome differences are still unbiased estimates of the program effect. Individuals further from the cut point on each side may be less similar to each other than those very near the cut point. The key to eliminating selection bias as data further from the cut point are used is to correctly model the relationship between the quantitative assignment variable and the outcome in the statistical analysis that controls for the influence on that outcome of differences on the assignment variable.

Key Concepts in Impact Evaluation

In the past decade, our understanding of impact evaluations and causal inference has increased substantially. In this section, we review some of the key concepts that have become important to a fuller understanding of estimating program effects. Although many of these concepts are also relevant to comparison group designs, their salience to the choices evaluators make and implement is clearer in the context of randomized designs and regression discontinuity studies.

Program Circumstances

One distinction often made in impact evaluation is between assessments of the efficacy of an intervention and those of its effectiveness, referred to respectively as **efficacy evaluation** and **effectiveness evaluation**. In this context, assessments of efficacy test an intervention under favorable circumstances, often in a relatively small study at a single site. These studies are frequently conducted by the developer of an intervention as an early “proof of concept” step for determining if it has promise for affecting the targeted outcomes. The delivery personnel for the intervention may be especially well trained (and may be the developers themselves), a high level of quality control may be applied to the service delivery, the participants may be selected to be especially appropriate, and the resources for supporting program delivery and client participation may be especially generous. Because establishing the efficacy of an intervention requires assurance that its effect estimates are valid, randomized designs are typically used. Those evaluations, however, are often conducted by the program developers themselves or others associated with the program development.

Assessments of effectiveness, in contrast, are oriented toward estimating the intervention effects for a fully deployed program implemented at scale and delivered as routine practice to typical members of the target population. Most ongoing programs studied by impact evaluators are of this sort. The circumstances of service delivery may be less than optimal, and participants will be typical for the program context, whether especially well matched to the program or not. The program developer or associated personnel may have provided training to the service delivery personnel, but they are not themselves part of the team that delivers the program. Depending on the situation, a randomized design may be desired that would usually be conducted by an independent evaluator; that is, one not affiliated with the program developer. Randomized designs used for effectiveness assessments are often referred to as randomized field experiments. Their purpose is to determine if the program has beneficial effects when implemented under real-world conditions of workaday practice.

Types of Counterfactuals

In all impact evaluations, program effect estimates are relative: They are estimated relative to the outcomes from whatever services the counterfactual group has access to or actually receives. In some randomized control trials for medical treatments, the control group does not receive any treatment. The treatment effect estimates are then relative to no treatment for the conditions the treatment is designed to address. However, it is more common for the control groups in impact evaluations to receive whatever program offerings or related services are available in the normal course of operations before or without access to the program being evaluated. For example, in an evaluation of a state-sponsored food assistance program, at least some members of the control group are likely to have access to food support provided by local charities, churches, and city governments.

For the impact evaluator, different counterfactual conditions answer different questions, and it is important to be clear on what the policy-relevant question is for the evaluation. Comparing program outcomes with conditions in which there are no organized interventions targeting those outcomes allows an estimate of the full inherent ability of the program to change those outcomes. This might be the focal interest for an efficacy study as described above. Or it may respond to the central policy question in a context in which there are, in fact, no other organized efforts targeting those outcomes.

However, there may be other services available to the target population, but the expectation of the program being evaluated is that it will add a component to the existing service system that will yield better overall effects. For example, mosquito nets for use while sleeping may be introduced in areas with a high incidence of malaria even though a range of mosquito abatement efforts are already under way in those areas. The policy-relevant question in that situation is not what the effects on malaria would be if there were no other mosquito control programs but, rather, whether the new net program adds to the effectiveness of what is already in place for the overall purpose of reducing the incidence of malaria. The counterfactual condition appropriate to that policy question is what is referred to as business as usual or practice as usual. The outcomes of

current efforts plus the program being evaluated are compared with those for current efforts without that program.

In still other situations the policy-relevant question may be whether the program being evaluated is more effective for improving the target outcomes than an existing program that might be replaced by the new program if it proves to be better. An impact evaluation of a promising new middle school math curriculum adopted in a school district might be such a situation. The school district already has a middle school math curriculum. The policy-relevant question is not how the new curriculum performs relative to no curriculum at all, or what the effects would be if the new curriculum were layered on top of the existing one. The question is simply whether it is better and should be preferred over the existing one. The appropriate counterfactual condition for the impact evaluation is then the current curriculum, with the evaluation comparing it head-to-head with the new curriculum being tried out in the evaluation. This too is a business-as-usual counterfactual, but with different implications for the conclusions that might be drawn from the impact evaluation results.

One aspect of business as usual as a policy-relevant counterfactual is that it is a dynamic rather than static basis for comparison. What this counterfactual condition consists of depends on the context and timing of the impact evaluation, and that can be different for the same program evaluated in different places or at different times. An example of this variability is the decrease in the program effect estimates that appeared in a series of evaluations of the Kindergarten Peer-Assisted Learning Strategies program conducted over a decade. As summarized in [Exhibit 8-A](#), the reason for the decreased effect estimates was not that the gains made by the program recipients had shrunk, but rather that the gains made by the control group increased over the years, apparently because of improvements in the business-as-usual conditions in the local schools.

Exhibit 8-A Changes in the Business-as-Usual Counterfactual Conditions: Five Randomized Control Evaluations of Kindergarten Peer-Assisted Learning Strategies

After comparing the results of five randomized control trials over a period of about 10 years with study samples drawn from the same community, the evaluators of the Kindergarten Peer-Assisted Learning Strategies (K-PALS) program found that the program effects had changed rather dramatically. The RCTs in the 1990s demonstrated

that low- and average-achieving students in the K-PALS program achieved statistically significant and educationally important improvements across a variety of early reading measures. But the effects had largely disappeared in two randomized control trials in 2004 and 2005. To investigate the mystery of the disappearing effects from this promising program, the evaluators examined the average gains made by the program and control groups in each of the five evaluations, with the results shown in the table below.

Outcome Measure	Study Year									
	1997		1998		1999		2004		2005	
	T	C	T	C	T	C	T	C	T	C
Rapid letter sounds	15.9	9.9	27.6	14.7	25.9	10.8	28.5	15.3	31.4	18.5
Segmenting	13.5	7.2	16.8	13.7	23.3	16.0	17.9	14.0	19.4	18.4
Word identification	6.1	3.7	7.9	5.7	8.8	5.0	12.7	10.3	14.2	17.4
Word attack	3.3	1.4	5.1	2.6	5.4	1.9	5.4	3.4	5.4	5.9

What this analysis revealed is that the gains from baseline to postintervention for program participants on all four outcomes were as large or larger in the later years as in the earlier ones. For instance, the kindergarteners exposed to the program showed gains of 6.1 points on the word identification measure in the 1997 study and 14.2 points in 2005. However, the gains for the business-as-usual control groups increased substantially over that period. On the word identification measure, the control group gains went from 3.7 points in 1997 to 17.4 points in 2005. The evaluators concluded that “the disappearing difference between treatment and control groups was likely because controls had improved their reading skills much more than they had in previous years” (Lemons, Fuchs, Gilbert, & Fuchs, 2014, p. 248). They speculated that this could be attributable to implementation of the federally required Reading First curriculum in kindergarten classes that used strategies similar to the K-PALS intervention.

Source: Adapted from Lemons, Fuchs, Gilbert, and Fuchs (2014).

Aside from the obvious importance of creating a control group that represents the counterfactual condition appropriate to addressing the policy-relevant questions for the impact evaluation, there is an ethical dimension to this issue. One of the objections to the use of randomized designs, with their inherent control of program access, that sometimes arises is the claim that needed services are being denied to control group participants—that something is being taken away from them. Note that none of the examples above of policy-relevant counterfactual conditions involve denying the control group access to services they would otherwise have if they were not in the control group. Indeed, it is difficult to imagine a circumstance in which a policy-relevant counterfactual would involve foreclosing

opportunities for a control group that were available to everyone else in a program's target population.

All counterfactual conditions, however, involve mutually exclusive options. A resident of a malaria-prone area either receives mosquito nets or not, and a middle school student experiences either the business-as-usual curriculum or the promising new curriculum. What is often meant by the claim that randomized impact evaluations deny opportunities to control groups is not that something is taken away that those groups already have, but rather that they do not have the opportunity to receive the program being evaluated—for instance, the mosquito nets or the new curriculum. That claim rather assumes that the benefits of the program being evaluated are already known or are so obvious that they do not need to be demonstrated. If that is actually the case, it would indeed be unethical to randomly assign individuals to receive or not receive that benefit. Randomized impact evaluations should be conducted only when there is uncertainty about the benefits of the program being evaluated, even the possibility that the program outcomes could be worse than current business as usual.

What complicates this issue is that program sponsors, providers, and advocates are generally quite convinced of the benefits of the program to which they have made such commitments, even though those benefits may not have been objectively demonstrated. This is a natural cognitive bias that may well be correct in some instances, but the history of impact evaluation is rife with examples in which such programs have proved to be no more effective than the business-as-usual alternative and, sometimes, less effective or even harmful. However, service providers may be so convinced that the program to be evaluated is effective that they are adamant that at least the neediest individuals must receive that program. This is a situation to which the regression discontinuity design is especially well suited and may be an acceptable alternative. A fuller discussion of the circumstances under which it is appropriate and ethical to conduct randomized impact evaluations is presented later in this chapter.

Types of Program Effects

Random assignment or assignment on the basis of a quantitative assignment variable in a regression discontinuity design sets up a contrast between a group offered access to the program being evaluated and a group not offered access to that program. In the ideal situation all those in the program group, and none of those in the control group, would actually participate in the program. That makes for a contrast that is aligned with the logic of the design and one that provides a clear interpretation of any differences in the outcomes between those two groups. This clean contrast is muddled if some individuals assigned to the program group do not actually participate in the program and/or some individuals assigned to the control condition nonetheless obtain the program services.

This situation, often labeled noncompliance with assignment or crossovers, led two pioneering evaluators to define and estimate **intent-to-treat (ITT) effects** in an evaluation of alternative police responses to domestic violence (Berk & Sherman, 1988). Intent-to-treat effect estimates compare outcomes for the individuals assigned to the program and control groups irrespective of whether those individuals actually complied with that assignment. This has the advantage of preserving the randomization or cut-point assignment that is the source of the rigor of the randomized and regression discontinuity designs. But intent-to-treat comparisons provide conservative program effect estimates when the crossovers dilute the outcomes for the program group and enhance those for the control group. In many circumstances, however, that comparison may be more relevant for policy because it takes into account the reality that not everyone in the target population with access to the program will actually participate in it. In that regard, intent-to-treat estimates may give the best indication of the net effects that can be expected if the program is offered at scale.

When the number of crossovers is relatively substantial, however, intent-to-treat comparisons do not answer another question program developers and other stakeholders often have: How effective is the program for those who fully experience it? Answering that question requires a comparison of outcomes for those who actually participated in the program with those who

did not participate irrespective of the condition to which the evaluation design assigned them. That comparison produces what are often called **treatment-on-the-treated (TOT) effects**. Note that it is usually TOT estimates that are generated in nonrandomized comparison group designs such as those described in the previous chapter. These typically begin with a group of individuals already participating in the program and compare their outcomes with a comparison group selected to have no program participation. As explained in that chapter, such comparisons are vulnerable to selection bias. Similarly, TOT program effect estimates derived from randomized and regression discontinuity designs have increased vulnerability to selection bias to the extent that they override the controlled assignment to conditions inherent in those designs.

Unit of Assignment

Our presentation so far has portrayed the controlled assignment to program and control conditions in randomized control designs and regression discontinuity designs mainly as one involving individuals. That is, individuals, whether persons or some other unit, are assigned to program and control conditions, access to the program is provided to the individuals in the program condition and not to the individuals in the control condition, and outcomes are measured on those individuals. The unit of assignment, whatever it is, is also the unit to which program access is offered or not, and is also the unit on which outcomes are measured. This could be a large aggregate unit, but it is the same unit in all aspects of the impact evaluation design. For example, a sample of communities might be randomly assigned to participate in an economic development program or not, with the program supporting community-level economic development initiatives, and such economic indicators as tax revenues and capital investments examined as outcomes.

There are useful variants of these designs, however, in which the unit of assignment to a program is an aggregate but, within an aggregate, the subunits experience either the program or control condition and each subunit's outcome is measured. The aggregate units in these designs are typically referred to as clusters. In a **cluster randomized trial**, for instance, clusters of individuals are randomly assigned to program and control conditions, and the individuals within each cluster either receive access to the program or not on the basis of the cluster assignment, and outcomes are measured on those individuals. This creates a multilevel design in which the units at the base level are described as being nested or clustered within the units at the higher level. Similar multilevel structures are possible for regression discontinuity designs and nonrandomized comparison group designs.

Multilevel designs of this sort can have advantages for impact evaluation. Aggregate units such as mental health agencies, daycare centers, social service offices, and schools can be recruited into the study and assigned to host the program being evaluated or continue with business as usual. The

individuals receiving services in those units can then be recruited to participate in the evaluation, but a representative sample within each unit may be sufficient and will reduce cost compared with data collection for everyone in the participating units. The cost of data collection may also be reduced because of the colocation of individuals within the participating units, thus limiting travel and related arrangements for data collectors. Additionally, because the individuals in the program and control conditions are in different sites, they and the associated program providers are unlikely to have the kind of routine interaction they would have if they were in the same sites. This reduces the potential for information about the program being evaluated to be shared with members of the control group in ways that would compromise the contrast between the conditions.

The advantages of cluster assignment to conditions nonetheless come with a downside. The individuals within each cluster are often more similar to one another than to individuals in other clusters. Patients served by the same mental health facility, for instance, will share characteristics associated with the catchment area for that facility as well as those related to their common experiences with the service of that facility. Such within-cluster similarities keep the outcome data for those patients from being statistically independent—there is some predictability from one to another on the basis of their shared membership in the cluster. Statistically dependent data require specialized analysis procedures. At the practical level, however, the main implication relates to the size of the sample needed. With cluster assignment, the number of individuals providing outcome data must be larger, possibly considerably larger, than the number required for individual random assignment in order to attain the same level of precision and statistical power to detect a program effect. The extent of the sample size inflation needed is determined by the number of clusters, how similar cluster members are to one another, and how dissimilar they are to individuals in other clusters. These matters, and the role of statistical power generally in impact evaluation, are discussed in the next chapter.

Multiple Intervention Conditions

To this point, the discussion has focused on a single program condition compared with a control condition. There is nothing about the controlled assignment designs discussed in this chapter or, for that matter, nonrandomized comparison group designs that restricts an impact evaluation to comparison of only two conditions. It may be desirable in some circumstances to include two or more different programs with similar goals in the evaluation, or variations on a particular program model, such as twice- versus once-weekly sessions. Multiple comparisons of this sort can be especially informative for policy and practice. For example, an international philanthropic organization concerned about teenage pregnancy may have some stakeholders who advocate school-based interventions with adolescents, while others advocate provision of long-acting reversible contraceptives without charge through local health clinics. To assess the effects of each of these options, and allow comparison of those effects with each other, the evaluators could recruit multiple sites and assign each to the school-based option, the health clinic option, or a business-as-usual control condition. This design requires recruiting more sites than required for an evaluation with only a single treatment arm, but fewer sites than would be needed for separate evaluations of each of the program options.

Multiple treatment arms need not involve different programs. A more common variant involves comparison of a larger versus a smaller dose or more and less intensive services. Comparisons of that sort can be especially informative for adjusting a program to be both effective and efficient. Examples include evaluations that compare half-day with full-day prekindergarten, or a 10-week substance abuse counseling program compared with one that lasts 20 weeks. Where the effectiveness of the service provided in one arm of a multiarm impact evaluation is already established, a business-as-usual control group may not be needed and, indeed, may even be considered unethical. In these cases, the evaluation may assign units only to treatment arms and omit the control condition. Comparative treatment effectiveness studies of this sort are increasingly common in fields like medicine, in which randomized clinical trials have already established the effectiveness of certain standard treatments so that

evaluation questions focus on whether promising innovative treatments can outperform those standard treatments.

When Is Random Assignment Ethical and Practical?

Most experts in impact evaluation and quantitative research methods consider randomized control designs to be the best choice for determining program effects because of the high level of internal validity for the effect estimates they produce when well executed. With no or minimal noncompliance with assignment and no or minimal attrition from outcome data collection, this design effectively eliminates selection bias and offers policymakers and other stakeholders the most methodologically credible estimates of average program effects possible with any impact design in the evaluation toolkit. However, there are many circumstances in which consideration of a randomized design for an impact evaluation raises ethical or practical issues that must be taken into account. Evaluators and other researchers who use randomized designs have been very thoughtful about these issues and have put forward various criteria with which to assess the appropriateness of a randomized design.

Ethical Considerations

An important set of criteria to justify the decision to use a randomized design takes the perspective of the potential benefits to society of the evaluation and protections of individual rights. For program evaluation, the potential benefits of a randomized control trial relate to the utility of the resulting quantitative estimate of the effects of a program on its target outcomes. For that estimate to have social benefit, it must be credible, but also actually valid and relatively unbiased, and be produced in a context in which it is likely to have some influence on decisions about the program. [Exhibit 8-B](#) summarizes the conditions under which a randomized design is justified that have been put forward by the Federal Judicial Center that carry this message authoritatively. The first of these conditions requires that the current situation be recognized as less than satisfactory, thus establishing the rationale for considering alternatives. The second condition specifies that the effectiveness of the alternative under consideration should be uncertain; for example, it may not have been tried or shown to be clearly effective in other jurisdictions. The third condition indicates that a randomized design should be the only practical means by which the effectiveness of the innovation at issue can be determined. A determination of effectiveness in this context means obtaining a credible program effect estimate. A less intrusive design, such as a comparison group design, thus is ruled out unless the practical circumstances allow it to provide an equally credible effect estimate. The fourth condition requires an a priori expectation that the results of the evaluation will influence decisions about whether to adopt the innovation under consideration. None of the first three conditions matter if there is no audience for the results of the evaluation with decision-making authority or influence.

Exhibit 8-B A Societal and Individual Protection Perspective on When to Randomize From the Federal Judicial Center

In 1978, Chief Justice Warren E. Burger, who served as the chairman of the board of the Federal Judicial Center, appointed the Advisory Committee on Experimentation in the Law. He charged the committee with studying the appropriateness and value of randomized experiments to evaluate innovations in the judicial system and making recommendations to guide the decision about when to use randomized experiments. [Table](#)

[8-B1](#) states the committee's five conditions for determining the appropriateness of using a randomized experiment.

Table 8-B1

TABLE 8-B1 Conditions for Deciding to Conduct Randomized Field Trials in the Justice Area

Federal Judicial Center	
1.	Present practice needs improvement or is of doubtful effectiveness.
2.	Significant uncertainty exists about the value of the proposed innovation.
3.	A randomized experiment is the only practical means of determining effectiveness of the innovation.
4.	Evidence to be used to inform future decisions about retaining status quo or implementing innovation.
5.	Experimentation may require different considerations, especially to protect human subjects in some fields, such as justice, than in others, such as medicine, because informed consent may not be possible.

Source: Federal Judicial Center, Advisory Committee on Experimentation in the Law (1981).

Source: Federal Judicial Center, Advisory Committee on Experimentation in the Law (1981).

The final condition is different in kind: It turns on the protection of human rights. The Federal Judicial Center's report acknowledges the difficulty of balancing the value of the evidence from a randomized control trial and the differential treatment of similar individuals inherent in that design. For example, if the innovation under consideration involves assigning individuals with similar criminal records and presenting offenses to different treatment options, it can raise questions about the ideal of equal treatment under the law. The *Belmont Report* (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979), which remains today a part of the guidance about the ethical treatment of human subjects in research studies for most federal agencies, established respect for persons as one of its three principles. Prisoners were specifically mentioned as a group that deserved special protection because of concern that they may be more vulnerable to coercion during recruitment of volunteers for randomized designs. The principle of respect for persons and their rights is especially relevant for controlled assignment designs, especially if the target population may not be in a position to freely give informed consent or have the capability to do so.

It is not unusual for some stakeholders to have ethical qualms about randomization, seeing it as arbitrarily and capriciously depriving control groups of positive benefits. The reasoning of such critics generally runs as follows: If it is worth evaluating a program with such an advanced approach as a randomized control trial (i.e., if the program seems likely to help participants), withholding that potentially helpful service from those who will be assigned to the control group is unethical. The counterargument is obvious: Ordinarily, it is not known whether an intervention is effective; indeed, that is the reason for the impact evaluation. Because researchers cannot know in advance whether an intervention will be helpful, they are not depriving the controls of something known to be beneficial and, indeed, may be sparing them from wasting time with an ineffective program. These concerns, however, reinforce the importance of directly addressing the degree of uncertainty about the benefits and potential harm of the program at issue before planning a randomized impact evaluation. Randomized designs excel at resolving that uncertainty, but are not appropriate if there is little uncertainty.

Practical Considerations

Another perspective on the question of when to randomize involves practical considerations. The technical and logistical resources required to mount and carry out a randomized control trial under field conditions are often substantial, though there are exceptions. A randomized impact evaluation thus should generally be undertaken only when there is sufficient prior reason to believe that the program to be evaluated has promising potential, or concerns that it may be harmful, and when it is methodologically feasible. Some of the steps that can be taken to assess these matters include the following:

1. Identify relevant prior studies and synthesize that literature to see if positive effects on the outcomes of interest were obtained with other interventions, or with less rigorous studies of the intervention at issue.
2. Pilot-test the intervention to establish its feasibility.
3. Examine the willingness of the target population to participate and adhere to the program regimen; review any evidence about how well the program is implemented.
4. Ensure that valid and reliable data collection instruments are available for the outcomes of interest.

When a randomized control design is both appropriate and feasible, some attention should be given to the nature of the program that will be evaluated. It is not unusual for the delivery of an intervention in a randomized evaluation to differ from how it is (or would be) delivered in routine practice. With standardized and easily delivered interventions, such as incentive payments, the experience provided to the intervention group in a randomized design is quite likely to be representative of the fully implemented program—there are only a limited number of ways that incentives can be delivered. More labor intensive, high skill interventions (e.g., job placement services, counseling, and teaching), on the other hand, may be delivered with greater care and consistency in a randomized control evaluation than when routinely provided by the program at scale. This phenomenon is known as the Hawthorne effect, so named for a classic study in which it became apparent that when participants knew they were

part of a research project, they behaved differently. Ideally, program providers and program recipients would be unaware of the fact that they were part of an evaluation study or, at least, unaware whether they were in the intervention or control group. That is very difficult to accomplish in the evaluation of social programs, however, especially randomized evaluations in which assignment to conditions generally requires consent and, even if not, is not easy to conceal.

In a similar vein, evaluators should be cautious about introducing any elements as part of the evaluation that may change the nature of the program. It is often appropriate and even necessary, for example, to provide incentives or tokens of appreciation to providers and participants for their cooperation with the evaluation. That makes the program being evaluated a combination of its intrinsic nature plus the atypical incentives. This may not matter if the incentives are modest, but could change the character of the program if they are more substantial.

Finally, we should note that the integrity of a randomized control trial is easily threatened. Although randomly formed intervention and control groups are expected to be statistically equivalent at the point of randomization, nonrandom processes may undermine that equivalence as the evaluation progresses. Differential attrition, for instance, may introduce differences between those intervention and control participants who do provide outcome data. Indeed, there are few, if any, large-scale randomized evaluations that have not been compromised to some extent by the inevitable departures from ideal circumstances. Even with such compromises, however, a randomized control trial will generally yield estimates of program effects that are more credible than any alternatives. [Exhibits 8-C](#) and [8-D](#) describe two evaluations in which randomized designs were implemented and illustrate many of the points raised above. The first of these relies on randomly assigning individuals to treatment and provides intent-to-treat estimates. The second uses a more complex evaluation involving cluster random assignment of schools to one of two program conditions and a control condition.

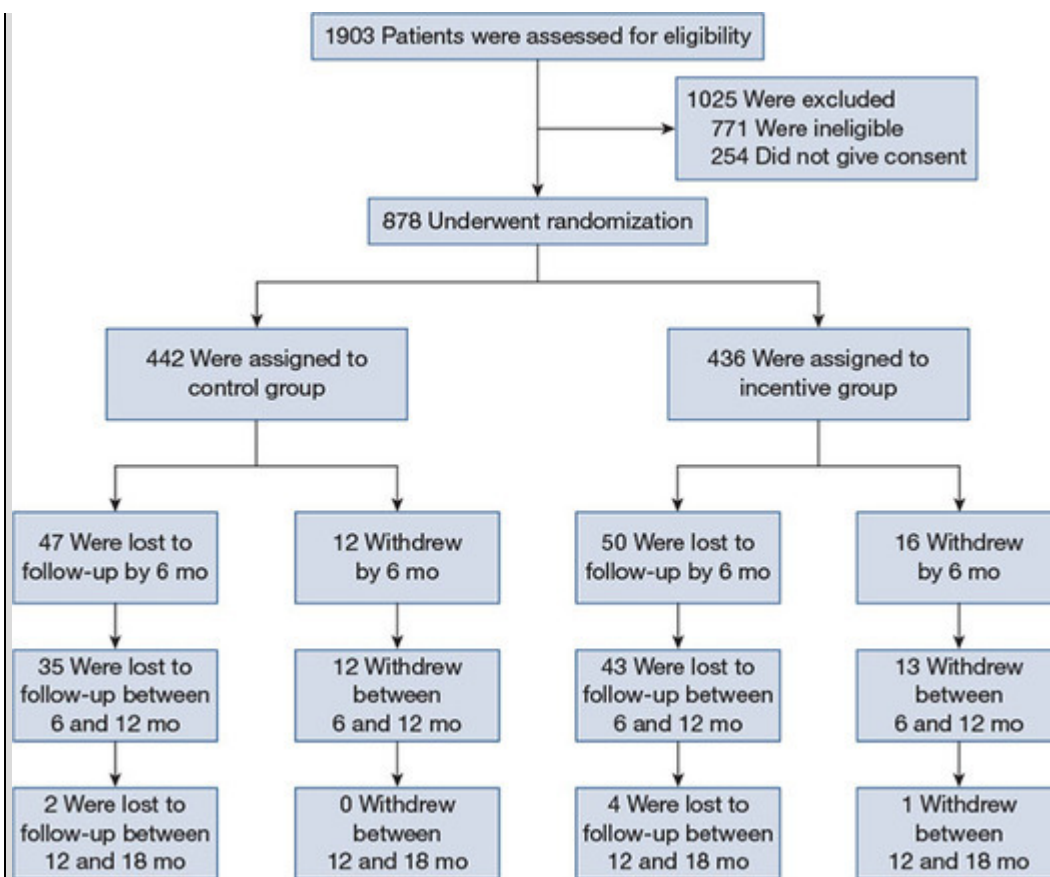
Application of the Regression Discontinuity Design

Regression discontinuity designs are appropriate in circumstances when the evaluator cannot randomly assign members of the target population into the treatment and control groups, but it is possible to assign members into these groups on the basis of their scores on a measure of their appropriateness for the intervention. Their appropriateness for the intervention might be based on need, merit, priority, or some other qualifying condition that can be used to divide the study sample or the entire target population into two groups. Those most appropriate by that standard are provided with access to the program, and those less appropriate do not receive access and are placed into the control group. The assignment into these two groups must be made on the basis of scores on a quantitative measure of the appropriateness, with a cut-point value on one side of which everyone receives access to the program and on the other side of which no one receives program access.

Exhibit 8-C A Randomized Control Evaluation of Financial Incentives for Smoking Cessation

To reduce smoking, which is the leading cause of preventable death in the United States, a major company offered financial incentives to encourage its employees to quit smoking. The incentives included \$100 for completion of a program aimed at assisting the employees' efforts to quit smoking, \$250 for complete cessation of smoking within 6 months after program enrollment, and \$400 for an additional 6 months without smoking. Eligibility for the incentives was based on being an adult smoker of more than five cigarettes per day who did not plan to leave the company in the next 18 months. All eligible employees who consented to being included in the evaluation were given information about community-based smoking cessation programs and the company's health insurance coverage for physician visits and prescriptions for smoking cessation treatment. This information provided a potential benefit to both the program and the control group, which could have been important for overcoming any objections to the randomization that determined who was also offered the cash incentives.

Assessment for Eligibility, Randomization, and Follow-Up



The figure below provides a breakdown of the study sample, beginning with recruitment, eligibility determination, and informed consent. Of 1,903 individuals initially recruited, 878 were randomly assigned to treatment (442) and control (436) conditions. The sample size was determined to be sufficient to detect a difference of 6.4% in the smoking cessation rates of the program and control groups after allowing for as much as 15% attrition from outcome measurement. Random assignment was done by first stratifying the sample by income level and amount of smoking (more than two packs per day or not), then assigning an equal number from each stratum to the program and the control conditions. By stratifying, which is also referred to as blocking, the evaluators could be confident that the two groups were similar on income and smoking history even if chance differences kept other characteristics from being totally equivalent. As expected, however, the randomization did result in the program and control groups being well balanced on a number of characteristics measured at the baseline.

Participants were interviewed 3 months after entering the study to determine if they had quit smoking and again at 6 months. A biochemical test was also administered to confirm participants' self-reports of complete cessation. All program effects were estimated in an intent-to-treat comparison on the basis of the original group assignment. In the table below, the effects of the program are shown, with 10.8% of the incentive program group completing a smoking cessation program compared with 2.5% of the control group. On the basis of smoking cessation reports confirmed by the biochemical test, 9.1% more of the incentive program group had quit smoking by 6 months and 9.7% more by the longer term checkup 6 months later. All of these differences were statistically significant, ruling out chance as a plausible explanation for the results. The authors summarize their

findings by saying, “This study shows that smoking cessation rates among company employees who were given both information about cessation programs and financial incentives to quit smoking were significantly higher than the rates among employees who were given program information but no financial incentives” (Volpp et al., 2009, p. 708).

Outcome Variable	Control Group (%) (n = 442)	Incentive Group (%) (n = 436)	Program Effect (Percentage Points)
Smoking cessation program			
Participation	5.4	15.4	10.0
Completion	2.5	10.8	8.3
Smoking cessation through 3 or 6 months			
Self-reported	14	23.4	9.4
Confirmed	11.8	20.9	9.1
Smoking cessation through 9 or 12 months			
Self-reported	6.1	15.1	9.0
Confirmed	5	14.7	9.7

Source: Volpp et al. (2009).

Source: Volpp et al. (2009).

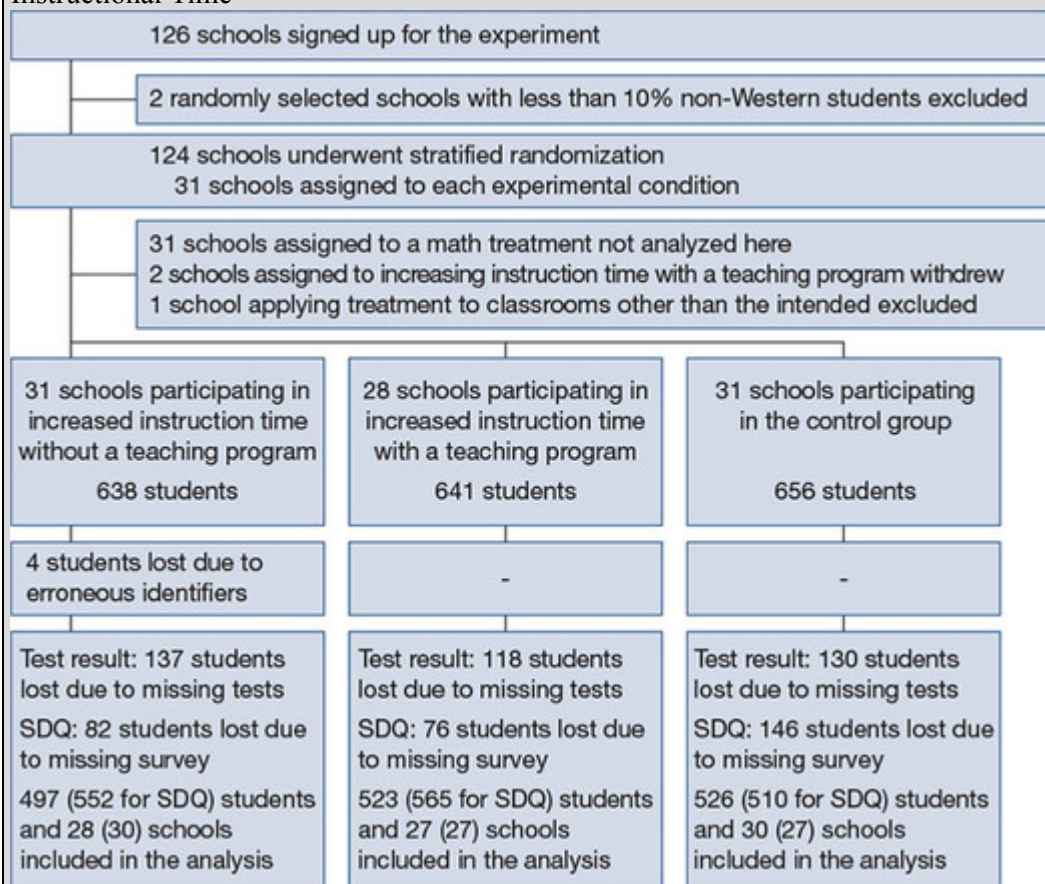
Exhibit 8-D A Cluster Randomized Control Evaluation of Increased Instructional Time for Reading

Although increasing instructional time seems like a logical solution to overcoming low levels of student achievement, there is relatively little evidence about its effectiveness. Also, there is a concern that additional time could backfire for students with lower levels of self-control. To evaluate the effects of increasing instructional time, the Danish Ministry of Education sponsored a cluster randomized field trial of the effects of expanding instructional time for reading, writing, and literature by 3 hours per week (15%) over 16 weeks. The evaluation was made more complex by including two different treatment arms. In one treatment arm, teachers were granted discretion in how to use the additional instructional time for reading. The stakeholders believed this would allow more individualized instruction. In the second arm, teachers were provided a detailed protocol for use of the instruction time developed by national experts. The outcome measures were (a) the Danish national reading exams covering language comprehension, decoding, and reading comprehension given to all fourth graders and (b) student responses to four subscales of the Strengths and Difficulties Questionnaire (emotional symptoms, conduct problems, peer relationship problems, and hyperactivity/inattention) that form a total behavioral difficulties index.

The Ministry of Education invited elementary schools with at least 10% non-native Danish speakers to participate in the evaluation, and 93 schools volunteered. Those schools were divided into blocks of 3 schools each that were matched on the percentage of students of non-Western origin and the average national reading test scores of the second graders in the prior year. The schools in each block were then randomly assigned to one of the two treatment groups or the business-as-usual control group. A single fourth grade classroom of students was selected at random from each school to contribute data. The baseline characteristics of the three groups were similar, including the students' prior test scores in reading and math. The figure at right provides a flowchart of the sample of schools and the students available for the evaluation and shows the amount of attrition, primarily because of missing test scores or surveys for the behavioral outcome.

The results from this evaluation demonstrated that increasing instructional time without a teaching protocol significantly increased overall reading scores and both the decoding and reading comprehension subscale scores. Increasing instructional time with a teaching protocol did not significantly increase overall reading scores but did increase the reading comprehension subscale scores, which was the focus of the teaching protocol used in that condition. The evaluation also found that the increased instructional time with a teaching protocol significantly decreased behavioral difficulties compared with the control group. In the treatment group without a teaching protocol, behavioral difficulties increased but not enough to be statistically significant.

The Assignment of Schools to the Danish Randomized Field Evaluation of Increased Instructional Time



The great advantage of this design for the impact evaluator is its inherent sense of fairness, combined with its ability when well executed to provide an unbiased program effect estimate. When resources are not sufficient to provide every member of the target population with program access or not all actually need the program, it appeals to many stakeholders' sense of fairness to provide access to those for whom the services are most appropriate. The design is flexible in that it allows the evaluator to collaborate with relevant program stakeholders to identify an appropriate study sample, which in some cases is the entire target population, and assign them to intervention and control conditions using criteria acceptable to those stakeholders.

The popularity of the regression discontinuity design has increased dramatically in recent years as impact evaluators have become more familiar with its advantages. Along with that popularity has come further development of the criteria that should be met to ensure valid program effect estimates. According to the standards set by one respected federal research unit, the U.S. Department of Education's What Works Clearinghouse, for example, the quantitative assignment variable must be at least ordinal (provide a sequence of values that range from low to high) and include a minimum of four or more unique values below the cut point and four or more unique values above the cut point (Deke et al., 2015).

One misconception sometimes arises when evaluators and stakeholders are considering a regression discontinuity design—that the assignment variable must be a valid measure of whatever it is that is to be the basis for assignment. For instance, if a measure of need for the program is to be used as the assignment variable, there may be concern about whether that variable really is an adequate measure of need. Stakeholders will understandably want that variable to direct those truly in the greatest need to the intervention group. The integrity of the regression discontinuity design, however, does not depend on having the assignment variable be a valid measure of anything. All that is required of the assignment variable is that it provide numerical scores along a continuum with a cut point that is strictly applied to determine program access.

Strict application of the cut point involves several different issues. For one, the cut-point score must be selected without any attempt to differentiate those somehow expected to have better outcomes with program exposure from those expected to have poorer outcomes. That could create selection bias, which would undermine one of the prime strengths of the regression discontinuity design. The cut point should be established on objective grounds, such as the number of units the program can serve, or an independently derived threshold for what is judged to constitute high need, risk for adverse outcomes, or the like as measured by the assignment variable.

It is also important that there be no manipulation of the values on the assignment variable that assigns some units to the program when their scores would not have made them eligible without that manipulation. For example, manipulation could occur when data from an intake form for mental health services provide a set of scores that are combined into a composite score used as the assignment variable to determine which patients will receive access to inpatient care. Suppose that clients with scores of 16 or above on a 20-point scale are to be assigned to inpatient care, while those below 16 are assigned to outpatient care (to test whether inpatient care produces better outcomes). Clinic staff members are likely to be aware of this cut point and during the intake process may form their own opinions about whether a new patient needs inpatient services. If they fudge the scoring on the intake form to push the composite score above the cut point for patients they believe need inpatient services, this would constitute manipulation.

A more blatant form of manipulation is simply to ignore the assignment made by the cut point and place the individual in the group deemed appropriate. This is especially tempting right around the cut point. Someone who does not understand the importance of strict application of the cut point may think that if an individual's score is really close on the control group side of the cut point, it shouldn't make any difference if that person is given access to the program anyway. These forms of manipulation can be identified by examining the proportions of individuals just above and just below the cut point to see if they are different. Manipulation would result in

more individuals than expected on one side of the cut point and fewer on the other.

Another crucial aspect of regression discontinuity designs is the attention that must be given to the statistical modeling that generates the program effect estimates. The individuals just above and just below the cut point are effectively randomized, which makes the difference in their outcomes a sound estimate of program effects. But in most applications there will be relatively few cases that close to the cut point, limiting the sample size if only those cases are analyzed and ignoring potentially useful information contributed by cases further from the cut point. Incorporating data from cases further from the cut point requires a statistical model that takes into account the underlying relationship between the assignment variable and the outcome measure. That underlying relationship represents a form of selection bias, but one completely and solely determined by the assignment variable. The appropriate statistical model can adjust for that built-in bias, making the resulting effect estimates unbiased.

Although practices differ among researchers using regression discontinuity designs, three statistical modeling approaches are most common. The traditional approach is to fit a regression model that predicts outcome scores with the assignment variable as a covariate used for statistical control (with other covariates possibly included as well) and a treatment variable that differentiates the intervention and control groups. Because the relationship between the assignment variable and the outcome scores may not be linear, this modeling approach typically includes higher order polynomial terms capable of accounting for various degrees of curvature in the relationship. And because the slope of the relationship, or any curvature, may not be the same on both sides of the cut point, the model typically also includes interaction terms that can account for that as well. It is common to find that some of these polynomial and interaction variables are not actually needed to account for the relationship between the assignment variable and the outcomes, and they may then be dropped from the model.

Another approach involves starting with a relatively narrow symmetric band of equal numbers of observations on each side of the cut point and using them to estimate the program effect with whatever statistical model

has been adopted. Additional bands are then progressively added, with the corresponding program effect estimated at each step. This process continues until program effect estimates that are qualitatively different from the first one are encountered. This allows the analyst to increase the sample size as much as possible without appreciably altering the effect estimate. As still another approach, evaluators may choose to overweight the observations closest to the cut point in their analysis and then progressively reduce the weights given to observations further from the cut point. Evaluators may provide estimates from more than one of these approaches to determine whether the program effect estimates are robust, that is, that they are not sensitive to the selection of a particular modeling approach.

Choosing an Impact Evaluation Design

The scientific credibility of well-executed randomized control designs for producing unbiased estimates of program effects would make them the obvious choice if such designs were typically the easiest, fastest, and least expensive to implement. Unfortunately, the control of program access via random assignment that is both the defining characteristic of randomized designs and the source of their rigor has a downside in the environment of social programs. The very idea of using the equivalent of a coin toss to determine who has access to a program that key stakeholders believe is beneficial, however undocumented that belief may be, is itself an obstacle to the use of randomized designs in impact evaluation. Even with a green light to implement random assignment, the practical challenges of recruiting a sample willing to participate in that process, organizing an uncompromised random assignment, and administering follow-up data collection activities under field conditions can be considerable.

Exhibit 8-E A Regression Discontinuity Evaluation of the Effects of Access to Health Insurance in Peru

Many developing countries have begun to provide public health insurance for those in poverty and without jobs in the formal economy that provide access to health insurance. Since late 2010 in Peru, individuals not working in the formal economy have been eligible for Social Health Insurance if they are among the lowest 25% on a welfare index known as the Household Targeting Index. Government officials calculate the index for each household from a household registry that is continuously updated and maintained, and which includes education of the head of household, type of materials used for flooring in the house, overcrowding of the dwelling, and other such variables. When eligibility is confirmed, insurance is made available at no cost to the eligible household members that provides broad coverage of health services from hospitals and health care centers operated by the Ministry of Health.

Bernal, Carpio, and Klein (2017) capitalized on the use of the Household Targeting Index to implement a regression discontinuity design to evaluate the short-term effects of this program. The requirement that households score in the lowest 25% on that index to be eligible for the insurance program provided an assignment variable and cut point that was already in place. Multiple variables collected on the household registry are used to calculate the index, and individuals do not know which are used for that purpose or their weights. It is thus unlikely that households were able to manipulate their scores on the index, so its integrity as an assignment variable was assumed to be high. Furthermore, when the researchers examined the proportion of the study population with values just

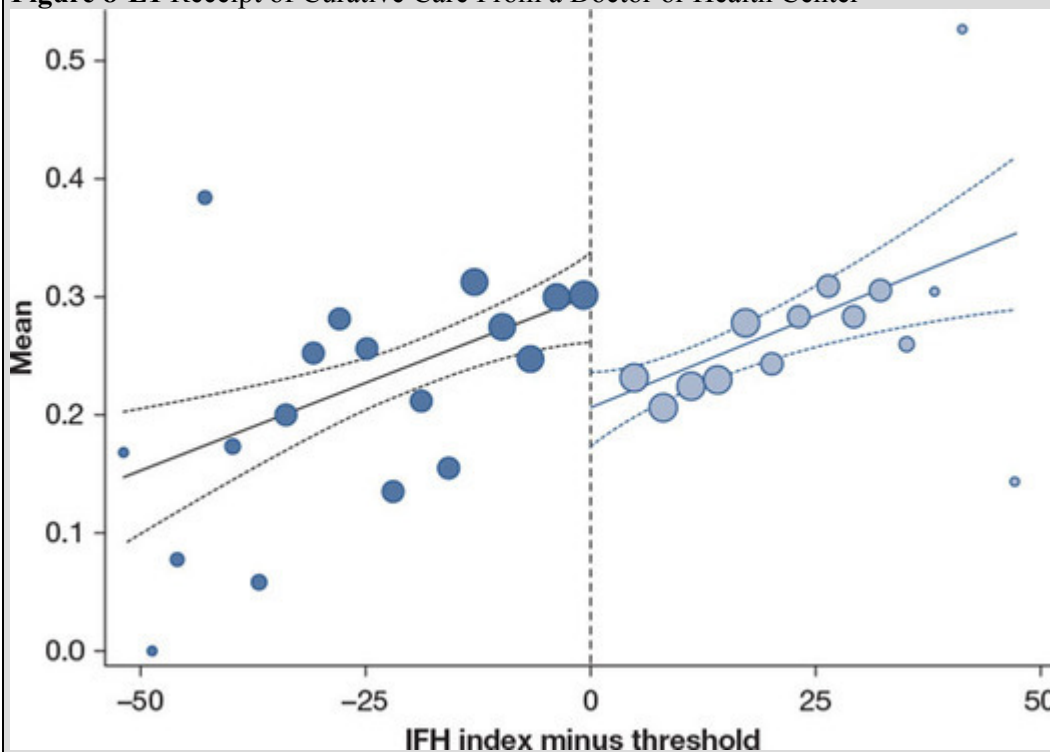
below the cut point, they found no evidence of the bunching of values that would appear if households had manipulated their scores in order to qualify for the program.

Outcome data were obtained from the National Household Survey of Peru, conducted in 2011, for a probability sample of 4,189 households with no formally employed adult in Lima Province, a densely populated area where there were numerous Ministry of Health facilities. Intent-to-treat program effects were estimated for those below the cut point on the Household Targeting Index and showed that individuals eligible for Social Health Insurance received more curative care (see [Figure 8-E1](#)), hospital and surgical care, medicines, and medical attention from a health care provider compared with those just above the cut point who were similar but ineligible for the program. Program effects estimated at various bandwidths closer to and further away from the cut point were found to be substantially similar.

The authors stated their conclusion this way: “We find strong effects of insurance coverage on arguably desirable, from a social welfare point of view, treatments such as visiting a hospital and receiving surgery and on forms of care that can be provided at relatively low cost, such as medical analysis in the first place and receiving medication” (Bernal et al., 2017, p. 134).

Source: Bernal, N., Carpio, M. & Klein, T. (2017). The effects of access to health insurance: Evidence from a regression discontinuity design in Peru. *Journal of Public Economics*, 154, 122-136. <https://doi.org/10.1016/j.jpubeco.2017.08.008>. Reprinted under the terms of a CC-BY 4.0 license: <https://creativecommons.org/licenses/by/4.0/>

Figure 8-E1 Receipt of Curative Care From a Doctor or Health Center



Although realistic, this is not a counsel of despair. Literally thousands of random assignment impact evaluations have been conducted, some under very challenging conditions. Moreover, they have made enormous contributions to knowledge about “what works” in the realm of social intervention. When there is uncertainty about a program’s effectiveness for improving the outcomes it targets, a rationale for the importance of having credible impact evidence, and a context within which there is a reasonable expectation that such evidence will have influence, the randomized control trial should be the design of choice. Evaluators should move to an alternative design only when there is good reason to believe that a randomized design is not appropriate to the situation or cannot be implemented with sufficient integrity.

The flexibility and rigor of the regression discontinuity design make it a good alternative choice when a randomized design has been ruled out. The quantitative assignment variable and its cut point that are the defining characteristics of this design can be adapted to select an intervention group in a way that may be both more acceptable and more feasible in the program context. That flexibility comes at some cost, however. Most notably, the highest quality estimates of the program effects come from the data clustered around the cut point. Depending on the nature of the assignment variable, the individuals in that narrow band may not be very representative of the entire intervention group. As the bandwidth is broadened to include more of the intervention and control groups, the validity of the effect estimates becomes more dependent on the adequacy of the statistical model used to generate those estimates. That produces technical challenges for the analysis and can result in a situation in which only the effect estimate in the narrowest range and based on the smallest subsample can be accepted with confidence. And sample size is a particular issue for the regression discontinuity design, which is rather greedy in this respect, requiring approximately two to three times the sample size as the comparable randomized design to have the same degree of statistical precision in the effect estimate.

The common characteristic of randomized and regression discontinuity designs is that strict control of access to the program (versus the control condition) is an inherent part of the design. None of the alternatives to these

two designs for impact evaluation have that characteristic. In some form or another, they are all based on a more or less natural sorting that produces conditions or groups of individuals with and without program exposure. There is no justification for assuming that an apples-to-apples comparison can be made under those circumstances that ensures that selection bias will not distort the program effect estimates.

The designs that lack control of program access and their limitations are discussed in some detail in [Chapter 7](#). Of those various designs, the most common is the nonrandomized comparison group design, in which outcomes are compared for a naturally occurring intervention group and a comparison group without program exposure that is assembled for that purpose. The value of these comparison group designs is that, when carefully done, they offer the prospect of providing plausible estimates of program effects while being relatively adaptable to circumstances where access to the program cannot be strictly controlled. Their advantages, however, rest entirely on their practicality and convenience in situations in which neither randomized designs nor regression discontinuity designs are feasible, not on their inherent rigor.

A critical question is how much risk for serious bias in estimating program effects there is when these nonrandomized comparison group designs are used. It is quite clear that poorly constructed versions of these designs are very vulnerable to bias, and that the magnitude of that bias can be considerable relative to the size of the actual program effects. The more relevant question is whether the risk for bias can be reduced to an acceptable level if these designs are well constructed and, if so, what it means for them to be well constructed. In recent years we have come closer to being able to answer these questions by drawing on a body of research that compares the results from comparison group designs with those from comparable randomized designs. Although these studies are becoming more common, the findings are still far from definitive. What the available work along these lines shows was reviewed in the previous chapter. In short, there are two procedures that are capable of reducing bias, and it appears that under favorable circumstances they may be sufficient to yield reasonably sound estimates of program effects. One of these involves drawing program and comparison samples that are similar in aggregate with

regard to their demographic mix, geographic location, and general social and cultural context. The other is effective use of well-chosen baseline covariates in the statistical analysis or matching. These covariates need to represent characteristics that are related to the outcome variables and on which the groups have consequential differences at baseline, and they need to include virtually all the independent characteristics with these properties.

The overall conclusion from the comparative evidence we have, therefore, indicates that, in a given application, impact evaluations using comparison group designs can yield effect estimates similar to those that would result from a randomized design or regression discontinuity design. But their ability to do so depends very much on the way they are constructed and implemented as well as the particular circumstances of the program and its participants. Furthermore, there is no direct test that can be applied to assess how valid the resulting effect estimates are, so the extent to which they are biased remains uncertain even under favorable conditions. Evaluators using designs without strict controls on access to the program, therefore, must rely heavily on a case-by-case analysis of the particular assumptions and requirements of the selected design and the specific characteristics of the program and target population to assess the likelihood that valid estimates of program effects will result.

A responsible evaluator faced with an impact evaluation opportunity has an obligation to carefully examine available alternative designs and advise stakeholders in advance about which of those alternatives seem feasible and their associated advantages and limitations. If the evaluation must proceed with something other than a randomized or regression discontinuity design, the evaluator should take special care to draw on all available resources, including the relevant research literature, in an effort to develop a design that will minimize the potential for bias. In reporting the findings of such an evaluation, the evaluator is also obligated to point out its limitations and the potential for bias despite whatever efforts have been made to minimize it.

Summary

- Impact evaluations are valued for their relevance to policy and practice, but will make misleading contributions if they misestimate program effects. The two impact evaluation designs with the greatest inherent ability to yield unbiased effect

estimates are *randomized control designs* and *regression discontinuity designs*. By controlling access to the program, these designs can eliminate selection bias and are therefore considered to be the most rigorous options available for impact evaluation.

- The distinctive feature of randomized control designs is random assignment of the relevant units to intervention and control groups. That procedure ensures that any initial differences between the groups occurs only by chance, and that their outcomes can be expected to be equal except for the effects of the program.
- Regression discontinuity designs control access to the program by assigning units to the intervention and control groups on the basis of whether their scores on a quantitative assignment variable are above or below a designated cut point. As the sole variable producing selection bias, once the influence of the assignment variable on the outcome is accounted for in an appropriate statistical model, this design can produce an unbiased estimate of the program effect in the region around the cut point.
- Randomized designs may raise ethical questions because of the way they control access to the program. A randomized design can be justified if the program addresses a condition recognized as unsatisfactory, the effectiveness of the program is uncertain, a randomized design is the best way to determine its effectiveness, the results will influence program decisions, and participants' rights will be protected.
- One distinction made in impact evaluation is between assessments of the *efficacy* of an intervention and assessments of its *effectiveness*. Assessments of efficacy ask about the effects of the program when it is implemented under relatively optimal circumstances, often as a proof-of-concept test. Assessments of effectiveness ask about the effects when the program is implemented as routine practice serving typical members of the target population.
- In impact evaluation, different counterfactual conditions answer different questions, and it is important to be clear about the policy-relevant question for the evaluation. Counterfactual conditions may involve no organized interventions targeting the same outcomes, or the business-as-usual support available in the absence of the program, or an alternative program with which the currently implemented program is compared.
- Randomized and regression discontinuity designs usually allow estimates of two kinds of program effects. *Intent-to-treat effect estimates* compare outcomes for the individuals assigned to the program and control groups irrespective of whether they actually complied with that assignment. *Treatment-on-the-treated estimates* compare outcomes for those who actually participated in the program with those who did not participate irrespective of the condition to which they were assigned.
- An evaluator asked to conduct an impact evaluation should carefully consider the advantages and limitations of alternative designs. Randomized designs have the greatest inherent capacity to produce unbiased program effect estimates, but may be difficult to implement for practical reasons. Regression discontinuity designs can also produce unbiased effect estimates and can be adapted to many evaluation circumstances, but not all. Nonrandomized comparison designs are often feasible and relatively easy to implement, but are the most vulnerable to bias.

Key Concepts

Assignment variable 188
Cluster randomized trial 195
Control group 186
Effectiveness evaluation 190
Efficacy evaluation 190
Intent-to-treat (ITT) effects 194
Quantitative assignment variable 188
Random assignment 187
Randomized control design 186
Regression discontinuity design 186
Treatment-on-the-treated (TOT) effects 194

Critical Thinking/Discussion Questions

1. Compare and contrast randomized designs and regression discontinuity designs. How do they differ in the way they attempt to minimize selection bias? How do they differ with regard to the demands they make on a program?
2. This chapter discusses four key concepts in impact evaluation. Describe those four key concepts and explain why each is important for impact evaluation.
3. What is the intent-to-treat effect? How is it related to the treatment-on-the-treated program effect estimate? What are the differences in the nature of the information provided by these two effect estimates?

Application Exercises

1. Locate a report of an impact evaluation that relied on random assignment. Summarize the evaluation design and discuss the practical issues involved in the application of random assignment in that evaluation.
2. Discuss the five ethical considerations for random assignment presented in the text. Propose a social intervention that would rely on random assignment and apply these five ethical principles. Would random assignment be ethical in evaluating that social intervention?

Chapter 9 Detecting, Interpreting, and Exploring Program Effects

[The Magnitude of a Program Effect](#)

[Detecting Program Effects](#)

[Practical Significance](#)

[Statistical Significance](#)

[Statistical Power](#)

[Examining Variation in Program Effects](#)

[Moderator Analysis](#)

[Mediator Analysis](#)

[The Role of Meta-Analysis](#)

[Informing an Impact Assessment](#)

[Informing the Evaluation Field](#)

[Summary.](#)

[Key Concepts](#)

The three previous chapters focused on the aspects of research designs for impact evaluation most relevant for obtaining valid estimates of program effects. In this chapter we first describe how the magnitude of program effects can be characterized, recognizing that some effects may be too small to be meaningful. This motivates a discussion of ways to assess the practical significance of program effects. It is essential that an impact evaluation be designed to detect at a statistically significant level any effect as large as or larger than the minimum judged to be of practical significance. This means that the research design must have adequate statistical power, and the factors that determine power and their implications for the evaluation design are discussed.

Although these considerations focus mainly on overall average program effects, the variability of effects can also be of interest. Two forms of analysis explore effect variability. Moderator analysis investigates differential effects for different participant subgroups. Mediator analysis investigates the causal pathways from proximal to distal outcomes by examining covariation in those outcomes. Finally, this chapter highlights the value to the impact evaluator of familiarity with prior evaluation research and notes the particular utility of meta-analyses that systematically synthesize such research. Aside from informing the practice of impact evaluation, meta-analysis is a vehicle for summarizing the growing body of knowledge about when, why, and for whom social programs are effective.

The end product of an impact evaluation is a set of estimates of the effects of the program on the outcomes measured. As discussed in [Chapters 6, 7, and 8](#), research designs vary in their vulnerability to various sources of bias, but if the resulting effect estimates are credible, they give some indication of the extent to which the program is effective. Interpreting the significance of those effect estimates, however, can be challenging, especially for stakeholders without a research background. In this chapter we describe the conventional ways in which the magnitude of a program effect is represented, how its practical significance can be characterized, and what is required to ensure that effects of practical significance are also statistically significant. We then discuss how the analysis of program effects can go beyond overall summary estimates to provide more differentiation about program effects for different subgroups in the target population and the causal pathways through which program effects are produced. At the end of the chapter, we briefly consider how meta-analyses that synthesize the effects found in multiple impact assessments can help improve the design and analysis of specific evaluations and contribute to the body of knowledge about social intervention.

The Magnitude of a Program Effect

The ability of an impact assessment to detect and describe program effects depends in large part on the magnitude of the effects the program produces. Small effects, of course, are more difficult to detect than larger ones, and their practical significance may also be more difficult to discern.

Understanding the issues involved in detecting and describing program effects requires that we first consider what is meant by the magnitude of a program effect.

In an impact evaluation, a program effect will appear as the difference between the outcome measured on the individuals (or other units) receiving the intervention and an estimate of what their outcome would have been had they not received the intervention. The most direct way to characterize the magnitude of the program effect, therefore, is simply as the numerical difference between the means of the two sets of outcome values. For example, a public health campaign might be aimed at persuading elderly persons at risk for hypertension to have their blood pressure tested. If a survey of the target population exposed to the campaign showed that the proportion tested during the past 6 months was .17, while the rate among seniors in a control condition was .12, the program effect would be a .05 increase in the rate. Similarly, if the mean score on a multi-item outcome measure of knowledge about hypertension was 34.5 for those exposed to the campaign and 27.2 for those in the control condition, the program effect on knowledge would be a gain of 7.3 points on that measure.

Characterizing the magnitude of a program effect in this manner can be useful for some purposes, but it is very specific to the particular measurement scale used to assess the outcome. Finding that knowledge of hypertension as measured on a multi-item questionnaire increased by 7.3 points among seniors exposed to the campaign will mean little to someone who is not very familiar with that questionnaire and how it is scored. To provide a general description of the magnitude of program effects, or to compare them statistically, it is usually more convenient and meaningful to represent them in a form that is not so closely tied to the specific measurement procedure.

One common way to indicate the general magnitude of a program effect is to describe it in terms of a percentage increase or decrease. For the campaign to get more seniors to take blood pressure tests, the increase in the rate from .12 to .17 represents a gain of 41.7% (calculated as $.05/.12$). The percentage by which a measured value has increased or decreased, however, is meaningful only for measures that have a true zero, that is, a point that represents a zero amount of the thing being measured. The rate at which seniors have their blood pressure checked would be .00 if none of them had done so within the 6-month period of interest. This is a true zero, and it is thus meaningful to describe the change as a 41.7% increase.

In contrast, the multi-item measure of knowledge about hypertension can only be scaled in arbitrary units. If the knowledge items were very difficult, a person could score zero on that instrument but still be reasonably knowledgeable; that is, not truly have zero knowledge. Seniors might, for instance, know a lot about hypertension but be unable to give an adequate definition of terms such as *systolic* and *calcium channel inhibitor*. In addition, the measurement scale might be constructed in such a manner that the lowest possible score was not zero but, maybe, 10. With this kind of scale, it would not be meaningful to describe the 7.3-point gain shown by the intervention group as a 27% increase in knowledge simply because 34.5 is numerically 27% greater than the control group score of 27.2. Had the scale been constructed and scored differently, the same actual difference in knowledge might have come out as a 10-point increase from a control group score of 75, which would yield a 13% change to describe exactly the same gain. When the scale of an outcome measure is in arbitrary units, the difference between the intervention and control groups on the measure will also be in arbitrary units, as will any representation of that difference as a percentage of any other arbitrary value on that measure.

Because many outcome measures are scaled in arbitrary units and lack a true zero, evaluators often use an **effect size statistic** to characterize the magnitude of a program effect rather than a raw difference score or simple percentage change. An effect size statistic expresses the magnitude of a program effect in a standardized form that makes it comparable across measures that use different units or scales.

The effect size statistic most commonly used to represent program effects that vary numerically, such as scores on a test, is the **standardized mean difference**. The standardized mean difference expresses the difference between the mean on the outcome measure for an intervention group and the mean for the control group in standard deviation units. The standard deviation is a statistical index of the variation across individuals or other units on a given measure that provides information about the range or spread of the scores. Describing the size of a program effect in standard deviation units, therefore, indicates how large it is relative to the variation in scores found within the respective intervention and control groups. Suppose, for example, that a test of reading readiness is used in an impact assessment of a preschool program, and that the mean score for the intervention group is half a standard deviation higher than that for the control group. In this case, the standardized mean difference effect size is .50. The utility of this effect size statistic is that it can be easily compared with, say, the standardized mean difference for a test of vocabulary that was calculated as .35. That comparison indicates that the preschool program was more effective in increasing reading readiness than vocabulary.

Some outcomes are binary rather than a matter of degree; that is, an individual either experiences some change or does not. Examples of binary outcomes include committing a delinquent act, becoming pregnant, or graduating from high school. For binary outcomes, an **odds ratio** effect size is often used to characterize the magnitude of a program effect. An odds ratio indicates how much smaller or larger the odds of an outcome event are for the intervention group compared with the control group. An odds ratio of 1.0 indicates even odds; that is, participants in the intervention group were no more and no less likely than controls to experience the change in question. Odds ratios greater than 1.0 indicate that intervention group members were more likely to experience a change; for instance, an odds ratio of 2.0 means that members of the intervention group were twice as likely to experience the outcome as members of the control group. Odds ratios smaller than 1.0 mean that they were less likely to do so. These two effect size statistics are described with examples in [Exhibit 9-A](#).

Standardized Mean Difference

The standardized mean difference effect size statistic is appropriate for representing intervention effects found on continuous outcome measures, that is, measures producing values that range over some continuum. Continuous measures include age, income, days of hospitalization, blood pressure readings, scores on achievement tests, and the like. The outcomes on such measures are typically presented in the form of mean values for the intervention and control groups, with the difference between those means indicating the size of the intervention effect. Correspondingly, the standardized mean difference effect size statistic is defined as

$$\frac{\bar{X}_i - \bar{X}_c}{sd_p},$$

where $\bar{X}_i$ is the mean score for the intervention group, $\bar{X}_c$ is the mean score for the control group, and sd_p is the pooled standard deviations of the intervention (sd_i) and

$$\sqrt{\frac{((n_i - 1)sd_i^2 + (n_c - 1)sd_c^2)}{(n_i + n_c - 2)}}$$

control (sd_c) group scores, specifically, with n_i and n_c the sample sizes of the intervention and control groups, respectively.

The standardized mean difference effect size, therefore, represents an intervention effect in standard deviation units. By convention, this effect size is given a positive value when the outcome is more favorable for the intervention group and a negative value if the control group is favored. For example, if the mean score on an environmental attitudes scale is 22.7 for an intervention group ($n_i = 25$, $sd_i = 4.8$) and 19.2 for the control group ($n_c = 20$, $sd_c = 4.5$), and higher scores represent a more positive outcome, the effect size would be

$$\frac{22.7 - 19.2}{\sqrt{((24)(4.8^2) + (19)(4.5^2)) / (25 + 20 - 2)}} = \frac{3.5}{4.7} = .74.$$

That is, the intervention group had attitudes toward the environment that were .74 standard deviations more positive than the control group on that outcome measure.

Odds Ratio

The odds ratio effect size statistic is designed to represent intervention effects on binary outcome measures, that is, measures with only two values such as arrested or not arrested, dead or alive, discharged or not discharged, pregnant or not, and the like. The outcomes on such measures are typically presented as the proportion of individuals in each of the two outcome categories for the intervention and control groups with one category viewed as a better outcome (success) and the other as a worse outcome (failure) in relation to the intended program effects. These data can be configured in a 2×2 table as follows:

	Success	Failure
Intervention group	p	$1 - p$
Control group	q	$1 - q$

where p is the proportion of individuals in the intervention group with a positive outcome, $1 - p$ is the proportion in the intervention with a negative outcome, q is the proportion of individuals in the control group with a positive outcome, and $1 - q$ is the proportion in the control group with a negative outcome; $p/(1 - p)$ is the odds of a positive outcome for an individual in the intervention group, and $q/(1 - q)$ is the odds of a positive outcome for an individual in the control group. The odds ratio is then defined as

$$\frac{p/(1-p)}{q/(1-q)}.$$

The odds ratio thus represents an intervention effect in terms of how much greater (or smaller) the odds of a positive outcome are for an individual in the intervention group than for an individual in the control group. For example, if 58% of the patients in a cognitive-behavioral program were no longer clinically depressed after treatment compared with 44% of those in the control group, the odds ratio would be

$$\frac{.58/.42}{.44/.56} = \frac{1.38}{.79} = 1.75.$$

Thus, the odds of being free of clinical levels of depression for those in the intervention group are 1.75 times greater than those for individuals in the control group.

Detecting Program Effects

The statistical representations of program effects found in impact evaluations, such as the effect size statistics described above, have a valence and a magnitude. *Valence* refers to the direction of the effect, algebraically represented by a plus or minus sign, but conceptually more appropriately viewed as indicating whether the intervention or control group had the more favorable outcome. Depending on the outcome measure, higher scores may be more favorable (e.g., income, achievement, health) or lower scores may be more favorable (e.g., unemployment, depression, mortality). The algebraic sign on the numerical difference between the mean outcome scores of the intervention and control groups, therefore, is not always aligned with the relevant valence on the effect size statistic. The magnitude of the statistical effect size, in turn, refers to how large it is numerically, a reflection of the size of the difference between the intervention and control group means on the respective outcome measures.

A systematic impact evaluation produces a statistical effect size estimate, but that observed effect size is not necessarily the true effect size, which is why it is characterized as an estimate. Aside from the potential for bias (discussed in previous chapters), there are always chance factors that contribute some amount of statistical noise to such estimates stemming from measurement error, the luck of the draw in selecting a research sample and dividing it into intervention and control groups, and other such sources of chance variation. Assessing whether the observed effect size is so large that it is unlikely to have resulted from such chance factors is the purpose of statistical significance testing. What it means to detect a program effect in an impact evaluation is that an appropriate statistical test indicates that the observed effect size is statistically significant and thus unlikely to have occurred simply by chance.

But there is no practical value in detecting a program effect that is trivially small, so small that it does not represent a worthwhile change in the relevant outcomes. Moreover, as will be evident in later discussion, it can be very challenging for the evaluator to design an impact evaluation capable of detecting very small effects. When designing an impact evaluation,

therefore, a critical step is specifying the smallest effect size that has practical significance in the context of the particular program, its objectives, and the outcome measures to be used. The impact evaluation must then be designed to detect effects that are as large as or larger than that minimum. In the context of impact evaluation, this is referred to as specifying the **minimum detectable effect size (MDES)** for which the evaluation will be designed.

Unfortunately, identifying an appropriate MDES is no simple matter. The numerical magnitude of an effect size statistic has no necessary relationship to the practical significance of that effect. A small statistical effect may represent a program effect of considerable practical significance, and a large statistical effect may be of little practical significance. For example, a very small reduction in the rate at which people with a particular illness must be hospitalized may have very important cost implications for health insurers. But improvements in their satisfaction with their care that are statistically larger may have negligible financial implications for those same stakeholders. The practical significance of statistical effect sizes can be assessed in various ways, some of the most useful of which we discuss next.

Practical Significance

Identifying the threshold at which a statistical effect size has practical significance in the context of an impact evaluation requires translation of statistical effect size metrics into terms relevant to the social conditions the program aims to improve. Sometimes this can be accomplished simply by restating the statistical effect size in terms of the outcome measure on which it is based, but only if that measure has readily interpretable practical significance. For juvenile delinquency programs, for instance, a common outcome measure is the rate of rearrest within a given time period after program participation. If a program reduces rearrest rates by 24%, this amount can readily be interpreted in terms of the number of juveniles affected and the number of delinquent offenses prevented. Among those familiar with juvenile delinquency, the practical significance of these effects is also readily interpretable. Effects on other inherently meaningful outcome measures, such as number of lives saved, amount of increase in annual income, and reduced rates of school dropouts, are likewise relatively easy to interpret with regard to their practical implications.

For many other outcome measures, bridging between statistical effect sizes and practical significance is not so easy. Consider a math tutoring program for low-performing sixth grade students with outcomes measured on a standardized mathematics test with scores that can range from 10 to 120, normed to have a standard deviation of 15. The statistical effect size is simply the difference in the mean scores of the intervention and control groups divided by 15 (e.g., a difference of 5 points would be an effect size of .33). But in practical terms, is a 5-point improvement in math skills on this test a big effect or a small one? Few people would be so intimately familiar with the items and scoring of this particular math achievement test that they could interpret statistical effects directly into practical terms.

Interpretation of statistical effects on outcome measures with values that are not inherently meaningful requires comparison with some external frame of reference that provides a practical context for those effects. With achievement tests, for instance, the average scores for students in different grades in the school might be available. Suppose that the mean score for

sixth graders in the school was 47 and the mean for seventh graders was 55. This 8-point increase (an effect size of .53, assuming a standard deviation of 15) thus represents the average increase in mathematics achievement scores associated with a full year of schooling. The evaluator and key stakeholders might agree that an effect of the math tutoring program that represents a 20% improvement over average grade level performance would be about the least they would expect from the program given the effort and cost it requires. The corresponding MDES for the impact evaluation thus would be .106 (20% of .53).

Some outcome measures may have a preestablished threshold value that can be used as a referent for interpreting the practical significance of statistical effects, or it may be possible to define a reasonable success threshold if one is not already defined. With such a threshold, statistical effects can be assessed in terms of the proportion of individuals above and below that threshold. For example, an impact evaluation of a mental health program that treats depression might plan to use the Beck Depression Inventory as an outcome measure. On this instrument, scores above 20 are generally recognized as indicating moderate to severe depression. One way to identify a minimal program effect that would have practical significance, therefore, is to ask the most relevant stakeholders to specify the smallest proportion of depressed patients moved below this threshold they would consider a worthwhile program effect.

Suppose in this example that intake data could be used to establish that 60% of the patients scored above the threshold for moderate to severe depression at baseline, and key stakeholders agreed that the least they would find acceptable is sufficient improvement in one fourth of those patients to move them below the threshold ($.25 \times .60 = .15$). This implies that the minimum acceptable change would increase the percentage below the threshold from 40% to 55%. These are referred to as a 40–60 and a 55–45 split in the under-over ratio of patients, respectively. Assuming a normal distribution of scores, a table of areas under the normal curve shows that a 40–60 split in the distribution occurs at a z score of $-.25$, and a 55–45 split occurs at a z score of $.13$. Z scores are in standard deviation units, so their difference of $.38$ provides the corresponding MDES value. Alternatively, with sufficient intake data the evaluator could convert the baseline scores into z scores

(subtracting the mean and dividing by the standard deviation) and make a similar calculation with the program data directly.

Another approach that can help evaluators and program stakeholders specify reasonable MDES values is to examine the distribution of effects found in evaluations of similar programs or programs with similar outcomes. In many program areas, meta-analyses of multiple studies have been conducted that analyze and report statistical effect sizes on relevant outcomes. For instance, a meta-analysis of the effects of marriage and relationship education programs (Hawkins, Blanchard, Baldwin, & Fawcett, 2008) reported that the mean effect size for the relationship quality outcome measures used in 46 well-controlled evaluation studies was .36. Though the range around this mean was not reported, an evaluator might judge that anything below, say, one fourth of that value (.09) was clearly a marginal performance for this kind of program on that outcome and select that value as the MDES.

In a policy context, an especially compelling approach to identifying an MDES that has practical significance for an impact evaluation is a cost-effectiveness analysis. Consider, for example, an outpatient treatment program for substance use disorders with an average total cost of \$5,000 per person treated. Assume further that a relapse within 2 years incurs an average total cost to public agencies of \$12,000 for the social workers, law enforcement personnel, further inpatient and outpatient treatment, and so on, that are involved in responding to a relapse. If the practice-as-usual treatment has a relapse rate of 60% (not atypical for addictive behaviors), the total treatment cost is \$500,000 for 100 patients and the relapse cost is \$720,000 (60 patients at \$12,000 each), for a total cost of \$1,220,000.

Policymakers might consider a somewhat more expensive (\$5,500 per person) innovative program a success if it could reduce the total cost by at least 10%. The treatment cost for 100 patients in that program is thus \$550,000 and the minimal target total cost is \$1,098,000 (a 10% reduction from \$1,220,000), so the alternative program would have to be effective enough to reduce the total relapse cost to no more than \$548,000 (so that the \$550,000 program cost plus \$548,000 relapse cost totals \$1,098,000). But some of the relapse cases will cycle back to the now more expensive

program, which is estimated to increase the 2-year relapse cost per patient from \$12,000 to \$12,175. To reduce the total relapse cost to \$548,000, the innovative program must then achieve a relapse rate of 45% (45 patients at \$12,175 each).

The usual effect size statistic for a binary outcome such as relapse (yes/no) is the odds ratio ([Exhibit 9-A](#)). The 2×2 table comparing a 60% relapse rate in the control group and a 45% relapse rate in the innovative treatment group looks like this:

	No Relapse	Relapse
Treatment	.55	.45
Control	.40	.60

The corresponding odds ratio computed to represent positive outcomes is $[(.55/.45) \div (.40/.60)] = 1.83$, which would then be the MDES that corresponds to the practical significance of the effects of the new more expensive program relative to the current program from the cost perspective of the policymakers.

There is no all-purpose best way to assess the practical implications of statistical effect sizes in order to identify an MDES with threshold practical significance, but approaches such as those we have just described will often be applicable and useful. An evaluator should be prepared to provide one or more translations of statistical effects into terms that can be more readily interpreted in the practical context within which the program operates. The particular form of the translation that will be most meaningful in any given context will vary, and the evaluator may need to be resourceful in developing a suitable interpretive framework. The approaches we have described and others that may be useful in some circumstances are itemized in [Exhibit 9-B](#), but this list does not exhaust the possibilities.

Exhibit 9-B Some Ways to Describe the Practical Significance of Statistical Effect Sizes

Difference on the Original Measurement Scale

When the original outcome measure has inherent practical meaning, the effect size may be stated directly as the difference between the outcome for the intervention and control groups on that measure. For example, the dollar value of health services used after a prevention program or the number of days of hospitalization after a program aimed at decreasing time to discharge would generally have inherent practical meaning in their respective contexts.

Comparison With Test Norms or Performance of a Normative Population

For programs that aim to raise the outcomes for a target population to mainstream levels, program effects may be stated in terms of the extent to which the program effect reduces the gap between the preintervention outcomes and the mainstream level. For example, the effects of a program for children who do not read well might be described in terms of how much closer their reading skills at outcome are to the norms for their grade level. Grade-level norms might come from the published test norms, or they might be determined by the reading scores of the other children who are in the same grade and school as the program participants.

Differences Between Criterion Groups

When data on relevant outcome measures are available for groups with recognized differences in the program context, program effects can be compared with those differences on the respective outcome measures. For instance, a mental health facility may use a depression scale at intake to distinguish between patients who can be treated on an outpatient basis and more severe cases that require inpatient treatment. Program effects measured on that depression scale could be compared with the difference between inpatient and outpatient intake scores to assess how they compare with that well-understood difference.

Proportion Over a Diagnostic or Other Preestablished Success Threshold

When a value on an outcome measure can be set as the threshold for success, the proportion of the intervention group with successful outcomes can be compared with the proportion of the control group with such outcomes. For example, the effects of an employment program on income might be expressed in terms of the proportion of the intervention group with household income above the federal poverty level in contrast to the proportion of the control group with income above that level.

Proportion Over an Arbitrary Success Threshold

Expressing a program effect in terms of a success rate may help depict its practical significance even if the success rate threshold is arbitrary. For example, the mean outcome value for the control group could be used as a threshold value. Roughly, 50% of the control group will be above that mean. The proportion of the intervention group above that same value will give some indication of the magnitude of the program effect. If, for instance, 55% of the intervention group is above the control group outcome mean, the program has not affected as many individuals as when 75% are above that mean.

Comparison With the Effects of Similar Programs

The evaluation literature may provide information about the statistical effects for similar programs on similar outcomes that can be compiled to identify those that are small and large relative to what other programs have achieved. Meta-analyses that systematically compile and report statistical effect sizes are especially useful for this purpose. An effect size for the number of consecutive days without smoking after a smoking cessation program could be viewed as having greater practical significance if it was above the average effect size reported in a meta-analysis of smoking cessation programs, and less practical significance if it was well below that average.

Conventional Guidelines

Cohen (1988) provided guidelines for what are generally “small,” “medium,” and “large” effect sizes in social science research. For standardized mean difference effect sizes, for instance, Cohen suggested that .20 was a small effect, .50 a medium one, and .80 a large one. However, these were put forward to illustrate the role of effect sizes in statistical power analysis, and Cohen cautioned against using them when the particular research context was known so that options more specific to that context were available. They are, nonetheless, widely used as rules of thumb for judging the magnitude of intervention effects despite their potentially misleading implications.

Statistical Significance

As noted above, what it means to detect an intervention effect in a systematic impact evaluation is that the effect estimate is statistically significant. No effect estimate can be assumed to be an exact estimate of the true effect. The outcome data on which an effect estimate is based always include some statistical noise that represents chance factors that create estimation error. Some chance factors, such as measurement error, influence the effect estimate directly, generally making the observed effect estimate smaller than it would be if that source of error were not inherent in the outcome data. The observed effect estimate is then further influenced by **sampling error**: the luck of the draw that produced the particular intervention and control samples of individuals (or other units) contributing data to the impact evaluation from the universe of samples that could have been selected. The primary determinant of sampling error is the size of the samples at issue; larger samples are less likely to differ from one another than smaller samples of the same target population.

The question statistical significance testing answers is whether an observed effect estimate is larger than is likely to have occurred merely as a result of sampling error that happened to yield samples with outcome values especially unrepresentative of the universe of possible samples from the target population. An estimate of the probability that sampling error could produce the observed effect estimate is derived by applying a statistical test appropriate to the sampling procedure used and the nature of the effect estimate tested. If that probability is less than a predetermined value (called alpha), which by convention is usually set at .05, the effect is deemed statistically significant.

Although the .05 alpha level has become conventional in the sense that it is used most frequently, there may be good reasons to use higher or lower levels in specific instances. When it is important for substantive reasons to have very high confidence in the judgment that a program is effective, the evaluator might set a more stringent threshold for accepting that judgment, say, an alpha level of .01. In other circumstances, for instance, exploratory

work seeking leads to promising interventions with modest sample sizes, the evaluator might use a less stringent threshold, such as an alpha of .10.

Statistical significance testing is thus the procedure an impact evaluator uses to determine if an acceptable claim can be made that a program effect has been detected. This is basically an all-or-nothing test. If the observed effect is statistically significant, it is at least minimally large enough to be discussed as a program effect. If it is not statistically significant, then no claim that it is a program effect will have credibility in the court of scientific opinion.

Statistical Power

With a specification of the smallest statistical effect size judged to have practical significance for a given program outcome (the MDES), a fundamental obligation for an impact evaluation is to be able to detect an effect that large or larger if the program actually produces an effect of that size. To ensure as much as possible that an MDES is detected, the impact evaluation must be designed to attain statistical significance if the program effect estimate is at least as large as the MDES. The statistical framework for developing that design revolves around the concept of **statistical power**. Statistical power is the probability that the estimate of the program effect will be found to be statistically significant if an effect of that size is determined, through an impact evaluation, to have occurred. The impact evaluator should design the evaluation design to have sufficient statistical power for the appropriate statistical significance test of the program effect estimate to reach statistical significance if that estimate represents an actual program effect as large as or larger than the specified MDES.

Statistical significance, recall, has to do with the probability that sampling error can be large enough to produce a nonzero effect estimate when there is no actual effect. Designing for statistical power in essence, then, is designing to keep sampling error sufficiently small relative to the magnitude of the actual underlying effect so that statistical significance will be attained when there is a real effect of that magnitude. There are four factors that determine statistical power: (a) the effect size to be detected, (b) the alpha level threshold for statistical significance, (c) the sample size, and (d) the statistical significance test used. The effect size to be detected in the context of impact evaluation is the MDES that has been identified to represent the threshold effect size for practical significance, so that is a given as a component of the statistical power function. The alpha level is set by the evaluator, by convention usually .05, and is thus also a given. The other two factors require further consideration.

Attaining such a high level of statistical power that it is near certainty that statistical significance will be achieved if the program produces an effect as large as the MDES is very difficult given the ever present chance of an

extreme sampling error fluke. The evaluator, therefore, must decide on an acceptable level of risk for what is called **Type II error** or beta error—failing to find statistical significance when there is in fact a real effect (the complement of **Type I error**—finding statistical significance when there is no actual effect—which is constrained by the alpha level set for significance testing). For instance, the evaluator could decide that the risk for failing to attain statistical significance for an actual effect at the MDES threshold level should be held to 10%; that is, $\beta = .10$.

Because statistical power is one minus the probability of Type II error, this means that the evaluator wants a research design that has a power of .90 for detecting an effect size at the selected threshold level or larger. Similarly, setting the risk for Type II error at .20 would correspond to a statistical power of .80. The latter is the conventional target for statistical power. Although not especially stringent for controlling Type II error on behalf of a potentially effective program, it is often realistic because of the practical difficulty of configuring the evaluation design to attain higher levels of power (e.g., a power of .95 that constrains the probability of Type II error to .05 or less).

What remains, then, is to design the impact evaluation with a sample size and appropriate statistical test that will yield the desired level of statistical power. The sample size factor is fairly straightforward: Larger samples increase the statistical power to detect an effect. Planning for the best statistical testing approach is not so straightforward. The most important consideration involves the use of baseline covariates in the statistical model applied in the analysis. Covariates were described in [Chapter 7](#) for use as control variables to adjust for baseline differences between intervention and comparison groups. In addition to that role, covariates correlated with the outcome measure also have the effect of extracting the associated variability in that outcome measure from the analysis of the program effect. Because statistical effect sizes involve ratios that are affected by the variance of the outcome measure (see [Exhibit 9-A](#)), these covariates inflate the representation of the statistical effect size in the analysis model and thus increase statistical power.

The most useful covariate for this purpose is generally the preintervention measure of the outcome variable itself. A pretest of this sort typically has a relatively large correlation with the later posttest and can thus greatly enhance statistical power, in addition to removing potential bias as described in [Chapter 7](#). To achieve this favorable result, the relevant covariates must be integrated into the analysis that assesses the statistical significance of the program effect estimate. The forms of statistical analysis that involve baseline covariates in this way include analysis of covariance, multiple regression, structural equation modeling, and repeated-measures analysis of variance.

It is beyond the scope of this text to discuss the technical details of statistical power estimation, sample size, and statistical analysis with and without covariates. Proficiency in these areas is critical for competent impact assessment, however, and should be represented on any evaluation team undertaking impact evaluations. More detailed information can be found in book-length treatments (e.g., Liu, 2014; Murphy, Myers, & Wolach, 2014), and a variety of computer programs are available to help with the necessary estimates and calculations.

Exhibit 9-C presents a representative illustration of the relationships among the factors that have the greatest influence on statistical power. It shows the sample sizes needed for various levels of statistical power to detect different MDES values with a typical statistical test of the difference between the intervention and control group means on an outcome variable with alpha set at the conventional .05 level. It also shows the great advantage of including a baseline covariate in the analysis that has a substantial correlation with the outcome measure, such as a pretest of that measure. Inclusion of a sufficiently strong covariate or group of covariates can reduce the required sample size by half or more for a given MDES and target statistical power.

Exhibit 9-C Interrelationships of Statistical Power, MDES, Baseline Covariates, and Sample Size

The practical difficulty of attaining adequate statistical power in an impact evaluation is greater with smaller MDES and higher levels of desired power. This is illustrated in the table below by showing the total sample size needed to achieve different power levels for a selection of MDES values. Also shown is the advantage of including a baseline covariate with a large correlation (.71) with the outcome measure in the analysis.

As revealed in this table, high levels of power for detecting small MDES values require quite large samples. Inclusion of a strong covariate greatly reduces the sample size needed, but it is still rather large for small MDES values. Many impact evaluations for social programs use total sample sizes of 500 or less (250 each in the intervention and control groups). The shading in the table distinguishes the samples smaller than 500. As is evident, with samples of 500 or less, high power can be attained only for relatively large MDES values (mainly $\geq .30$), despite the fact that smaller MDES values will have practical significance for the primary outcomes of many programs.

Power (beta)	Covariate	MDES				
		.10	.20	.30	.40	.50
.70 (.30)	No	2,471	619	276	156	101
	Yes	1,237	311	139	79	52
.80 (.20)	No	3,142	787	351	198	128
	Yes	1,572	395	177	100	65
.90 (.10)	No	4,205	1,053	469	265	170
	Yes	2,104	527	236	133	86
.95 (.05)	No	5,200	1,302	580	327	210
	Yes	2,601	652	291	165	106
.99 (.01)	No	7,352	1,840	820	462	297
	Yes	3,677	922	411	233	150

Note: Alpha = .05. MDES represented as the standardized mean difference effect size. Total sample size divided evenly between intervention and control groups. Baseline covariate that correlates .71 with the outcome measure, accounting for 50% of the variance on that measure. Power calculations done with PowerUp! software (Dong & Maynard, 2013; Google "PowerUp! software" to locate current source for free download).

Note: Alpha = .05. MDES represented as the standardized mean difference effect size. Total sample size divided evenly between intervention and control groups. Baseline covariate that correlates .71 with the outcome measure, accounting for 50% of the variance on that measure. Power calculations done with PowerUp! software (Dong & Maynard, 2013; Google "PowerUp! software" to locate current source for free download).

Close examination of the table in [Exhibit 9-C](#) will reveal how difficult it can be to achieve adequate statistical power in an impact evaluation. High power is attained only when either the sample size or the MDES is rather large. Both of these conditions are often unrealistic for impact evaluation. Suppose, for instance, that the evaluator decides to hold the risk for Type II error to the same 5% level customary for Type I error ($\beta = \alpha = .05$),

corresponding to a .95 power level. This is a quite reasonable objective in light of the unjustified damage that might be done to a program if it produces meaningful effects that the impact evaluation fails to detect at a statistically significant level. Suppose, further, that the evaluator determines that an MDES of .20 on the outcome at issue would represent a positive program accomplishment and therefore should be detected. Table 9-C1 indicates that achieving that much statistical power would require a total sample of 1,302 individuals, 651 in each group (intervention and control). Including a strong covariate reduces the required total sample appreciably to 652 (326 in each group). Although such numbers may be attainable in some evaluation situations, they are far larger than the sample sizes reported in many impact evaluations. The sample size demands are even greater if the relevant MDES is below .20, which is not unrealistic for the primary outcomes of many social programs.

The statistical power demands are even greater for the multilevel impact evaluation designs described in [Chapter 8](#) in which the unit of assignment is a cluster with subunits that provide the outcome data. As noted in that chapter, these designs have distinct advantages in some situations and are increasingly common. Cluster randomized designs, for instance, are often used in educational evaluations with schools or classroom assigned to intervention and control conditions and outcomes measured on the students within those clusters. Attaining adequate statistical power is an especial challenge in such multilevel designs because the individuals within clusters are likely to be more similar to one another than to individuals in other clusters. For statistical purposes, that similarity means that the information contributed by each individual to the outcome data is somewhat redundant with that contributed by the other individuals in the cluster, a situation known as statistical dependency. The result is that the **effective sample size** that counts toward statistical power is smaller than the actual total number of individuals in all the clusters.

When there is more similarity among the individuals within clusters, there will be correspondingly less similarity across the clusters. A statistic called the intraclass correlation coefficient (ICC) is used to represent the between-cluster variation on the outcome as a proportion of the total variance (between- plus within-cluster variance). For a given total sample size, the

effective sample size and hence statistical power are reduced as the ICC increases above zero. And for a given total sample size and a given ICC, statistical power is increased as the number of clusters increases (more clusters come closer to individual-level assignment where there is no cluster effect). In [Exhibit 9-D](#), we show these statistical power patterns for a total sample of 1,000 individuals divided into different numbers of clusters with different ICC values. Although the power to detect an MDES with individual level assignment (no clusters or, one might say, 1,000 clusters of 1 person each) is quite high (.98), it drops quite rapidly as the number of clusters decreases and the ICC increases. Especially notable is the considerable deterioration in statistical power with ICC values as small as .01 and .05.

The illustrative statistical power results in Tables 9-C1 and 9-D1 are rather sobering from the perspective of impact evaluation. Many of the scenarios depicted there demonstrate the practical difficulty of achieving a high level of statistical power for modest MDES values with the sample sizes available in many evaluation situations. It is not unusual for MDES values in the range of .10 to .30 to represent program effects large enough to have practical significance. When impact evaluations are underpowered for such effects, there is a larger than desired probability that they will not be statistically significant despite their practical significance. That result is generally interpreted as a failure of the program to produce effects, which is not only technically incorrect but quite unfair to the program administrators and staff. Such findings mean only that the effect estimates are not reliably larger than sampling error, which itself is large in an underpowered study, not that they are necessarily small or zero. These nuances, however, are not likely to offset the impression of failure given by a report for an impact evaluation that found no statistically significant effects.

Exhibit 9-D The Implications of Cluster Assignment for Statistical Power

In recent years, many impact evaluations have departed from the assignment of individuals to intervention and control groups in circumstances in which that presents practical difficulties and, instead, have assigned the groupings or clusters in which those individuals are embedded (e.g., mental health facilities with their associated patients).

The cost of choosing cluster assignment is mainly in the reduction of statistical power compared with individual-level assignment when the sample size is the same for both.

The extent of that reduction in power will depend on the number of clusters and the similarity of the members within clusters relative to the similarity across clusters, the latter indexed by a statistic called the intraclass correlation coefficient (ICC).

In the table below, we show the statistical power for various scenarios that differ in the number of clusters that are assigned and the ICCs for those clusters. In all these scenarios the MDES is .25, the total sample size is 1,000, and significance is tested with $\alpha = .05$.

Total sample = 1,000 MDES = .25	ICC					
Number of Clusters	.00	.01	.05	.10	.20	.30
10 (100 per cluster)	.90	.68	.24	.14	.08	.07
20 (50 per cluster)	.96	.86	.51	.32	.19	.14
50 (20 per cluster)	.97	.94	.79	.62	.42	.31
1,000 (1 per cluster) [no clustering]	.98					

Note: Total sample size of 1,000 evenly divided between the intervention and control groups; MDES of .25. Outcomes are measured at the individual level. Statistical significance is tested at $\alpha = .05$ (two-tailed). No baseline covariates are included in the analysis model. Power calculations were done with PowerUp! software (Dong & Maynard, 2013; Google "PowerUp! software" to locate current source for free download).

Note: Total sample size of 1,000 evenly divided between the intervention and control groups; MDES of .25. Outcomes are measured at the individual level. Statistical significance is tested at $\alpha = .05$ (two-tailed). No baseline covariates are included in the analysis model. Power calculations were done with PowerUp! software (Dong & Maynard, 2013; Google "PowerUp! software" to locate current source for free download).

Reading across the rows in Table 9-D1 reveals how rapidly statistical power declines with increasing ICC values, including even with the smallest values. Reading down the columns shows the increase in statistical power associated with more clusters, each with fewer individuals. At the extreme, there are as many clusters as individuals, which means individual-level assignment, and the ICC is necessarily zero and power is at a maximum for this total sample size.

Examining Variation in Program Effects

So far, our discussion of program effects has focused on the overall mean effects on relevant outcome measures. However, program effects are rarely identical for all the subgroups in a target population or for all outcomes, and the variation in effects should also be of interest to an evaluator. Examining such variation requires that other variables be brought into the picture in addition to the outcome measure of primary interest and covariates. When attention is directed toward possible differences in program effects for subgroups of the target population, the additional variables define the subgroups to be analyzed and are called moderator variables. For examining how varying program effects on one outcome variable are related to the effects on another outcome variable, both outcome variables must be included in the analysis with one of them tested as a potential mediator variable. The sections that follow describe how variations in program effects can be related to moderator or mediator variables and how the evaluator can uncover those relationships to better understand the nature of the program's impact on the target population.

Moderator Analysis

A **moderator variable** characterizes subgroups in an impact assessment for which the program effects may differ. For instance, gender would be such a variable when considering whether a program effect was different for males and females. To explore this possibility, the evaluator could divide both the intervention and control groups into male and female subgroups, determine the mean program effect on a particular outcome for each gender, and then compare those effects. An alternative approach that makes more efficient use of the data is to use the moderator variable in an interaction term entered into a multiple regression analysis predicting the outcome variable from treatment status (intervention vs. control group). The pertinent interaction term consists of the cross-product of the moderator variable and the treatment variable. Construction of interaction terms is described in any text on multiple regression analysis.

It is relatively common for moderator analysis to reveal variations in program effects for different subgroups. The major demographic variables of gender, age, race/ethnicity, and socioeconomic status often characterize groups that respond differently to a social program. It is, of course, useful for program stakeholders to know which groups benefit the most and least from the program as well as the overall average effects. For example, focusing attention on the groups receiving the least benefit from the program and finding ways to boost the effects for those groups is an obvious way to strengthen a program and increase its overall effectiveness. The investigation of moderator variables, therefore, is often an important aspect of impact assessment. Those analyses may identify subgroups for which program effects are especially large or small, reveal program effects for some types of participants even when the overall mean program effect is small, and allow the evaluator to probe the outcome data in ways that strengthen the overall conclusions about the program's effectiveness.

Evaluators can most confidently and clearly detect variations in program effects for different subgroups when the subgroups can be defined at the start of the impact assessment. In that case, there are no selection biases involved. For example, a participant in the evaluation study obviously does

not become a male or a female as a result of selection processes at work during the period of the intervention. However, selection biases can come into play when subgroups are defined that emerge or change during the course of the intervention. For example, if some members of the control and intervention groups move away after the study is under way, then the program may have influenced these behaviors, and the subgroup analysis could be biased. Consequently, the analysis needs to take into account any selection biases in the formation of such emergent subgroups.

If the evaluator has measured relevant moderator variables, it can be particularly informative to examine differential program effects for those individuals most in need of the benefits the program attempts to provide. It is not unusual to find that program effects are smallest for those most in need at the time when they were recruited into the evaluation study. An employment training program, for instance, will typically show better job placement outcomes for participants with recent employment experience and some job-related skills than for chronically unemployed persons with little experience and few skills. Although that itself is not surprising or necessarily a flaw in the program, moderator analysis can reveal whether the neediest cases receive any benefit at all. If positive program effects appear only for the less needy and are trivial for those most in need, the implications for improving the program are quite different than if the neediest benefit, but by a smaller amount.

The differential effects of the employment training program in this example could be so strong that the overall average effect of the program on, say, later earnings might be large despite a null effect on the chronically unemployed subgroup. Without moderator analysis, the overall positive effect would mask the fact that the program was ineffective with a critical subgroup. Such masking can work the other way as well. The overall program effect may be negligible, suggesting that the program was ineffective. Moderator analysis, however, may reveal large effects for a particular subgroup that are washed out in the overall results by poor outcomes in larger groups. This can happen easily, for instance, with programs that provide universal services that cover individuals who are not at risk for the behavior or other outcome the program is attempting to influence, the “bulletproof” subgroup mentioned in [Chapter 6](#). A broad drug

prevention program in a middle school, for example, will involve many students who do not use drugs and have little risk of ever using them. No matter how good the program is, it cannot improve on the drug-use outcomes these students will have anyway. An important test of program effects would thus be a moderator analysis examining outcomes for the subgroup that is at high risk.

One important role of moderator analysis, therefore, is to avoid premature conclusions about program effectiveness based only on the overall average program effects. A program with overall positive effects may still not be effective with all types of participants. Similarly, a program that shows no overall effects may be quite effective with some subgroups. Another possibility that is rare but especially important to diagnose with moderator analysis is a mix of positive and negative effects. A program could have systematically harmful effects on one subgroup of participants that could be masked in the overall effect estimates by positive effects for other subgroups. A program that works with juvenile delinquents in a group format, for instance, might successfully reduce the subsequent delinquency of the more serious offenders in the group. The less serious offenders in the mix, on the other hand, may be more influenced by their peer relations with the serious offenders than by the program and actually increase their delinquency rates. Depending on the proportions of more and less serious offenders, this negative effect may not be evident in the overall mean effect on delinquency for the whole program.

In addition to uncovering differential program effects, evaluators can use moderator analysis to test their expectations about what differential effects should appear. This can be especially helpful for probing the consistency of the findings of an impact evaluation and strengthening the overall conclusions about program effects that are drawn from those findings. [Chapters 6, 7, and 8](#) discuss the many possible sources of bias and ambiguity that can complicate attempts to derive a valid estimate of program effects. Although there is no good substitute for methodologically sound measurement and design, selective probing of the patterns of differential program effects can provide another check on the plausibility that the program itself has produced the effects observed and not some uncontrolled influence on the outcomes.

One form of useful probing, for instance, is dose-response analysis for participants in the intervention group. This concept derives from medical research and reflects the expectation that, other things equal, a larger dose of the treatment should produce more benefit, at least up to some optimal dose level. It is difficult to keep all other things equal, of course, but it is still generally informative for the evaluator to conduct moderator analysis when possible on the outcomes associated with differential amount, quality, or type of service. Such analysis is especially informative when it can be applied to distinct groups of study participants with different program experiences. Suppose, for instance, that a program has two service delivery sites that serve a similar clientele, each with intervention and control groups. If the program has been more fully implemented at one site than the other, the program effects would be expected to be larger at that site. If they are not, and especially if they are larger at the weaker site, this inconsistency casts doubt on the presumption that the effects being measured actually stem from the program and not from other sources. Of course, there may be a reasonable explanation for this apparent inconsistency, such as faulty implementation data or an unrecognized difference in the nature of the clientele, but the analysis still has the potential to alert the evaluator to possible problems in the logic supporting the conclusions of the impact assessment.

Another example in a similar spirit comes from a classic impact evaluation, the time series study of the effects of the British Breathalyzer crackdown on traffic accidents (Ross, Campbell, & Glass, 1970). Because the time series design used in that evaluation is not especially strong for isolating program effects, an important part of the evaluation was a moderator analysis that examined differential effects for weekday commuting hours in comparison with weekend nights. The researchers' expectation was that if it was the program that produced the observed reduction in accidents, the effects should be larger during the weekend nights, when drinking and driving were more likely, than during daytime commuter hours. The results confirmed that expectation and thus lent support to the conclusion that the program was effective. Had the results turned out the other way, with a larger effect during commuter hours, the plausibility of that conclusion would have been greatly weakened.

The logic of moderator analysis aimed at probing conclusions about the program's role in producing the observed effects is thus one of checking whether expectations about differential effects are confirmed. The evaluator reasons that if the program is operating as expected and truly having effects, those effects should be larger here and smaller there, for example, larger where the behavior targeted for change is most prevalent, where more or better service is delivered, for groups that should be naturally more responsive, and so forth. If appropriate moderator analysis confirms these expectations, it provides supporting evidence about the existence of program effects. Most notably, if such analysis fails to confirm straightforward expectations, it serves as a caution to the evaluator that there may be influences on the program effect estimates other than the program.

While recognizing the value of moderator analysis for probing the plausibility of conclusions about program effects, we must also point out the hazards. For example, the amount of services or dose received by program participants is not randomly assigned, so comparisons among subgroups on moderator variables related to amount of service may be biased. The program participants who receive the most service, for instance, may be those with the most serious problems. For example, a school reform program may use coaches to improve teachers' instructional practices. If the teachers who struggle in the classroom receive more coaching, then a simple dose-response analysis will likely show smaller effects for those receiving the larger doses of service. However, if the evaluator looks at dose differences within groups with similar levels of need, the expected dose-response relation may appear. Clearly, there are limits to the interpretability of moderator analysis aimed at testing for program effects, which is why we present it as a supplement to good impact evaluation design, not as a substitute.

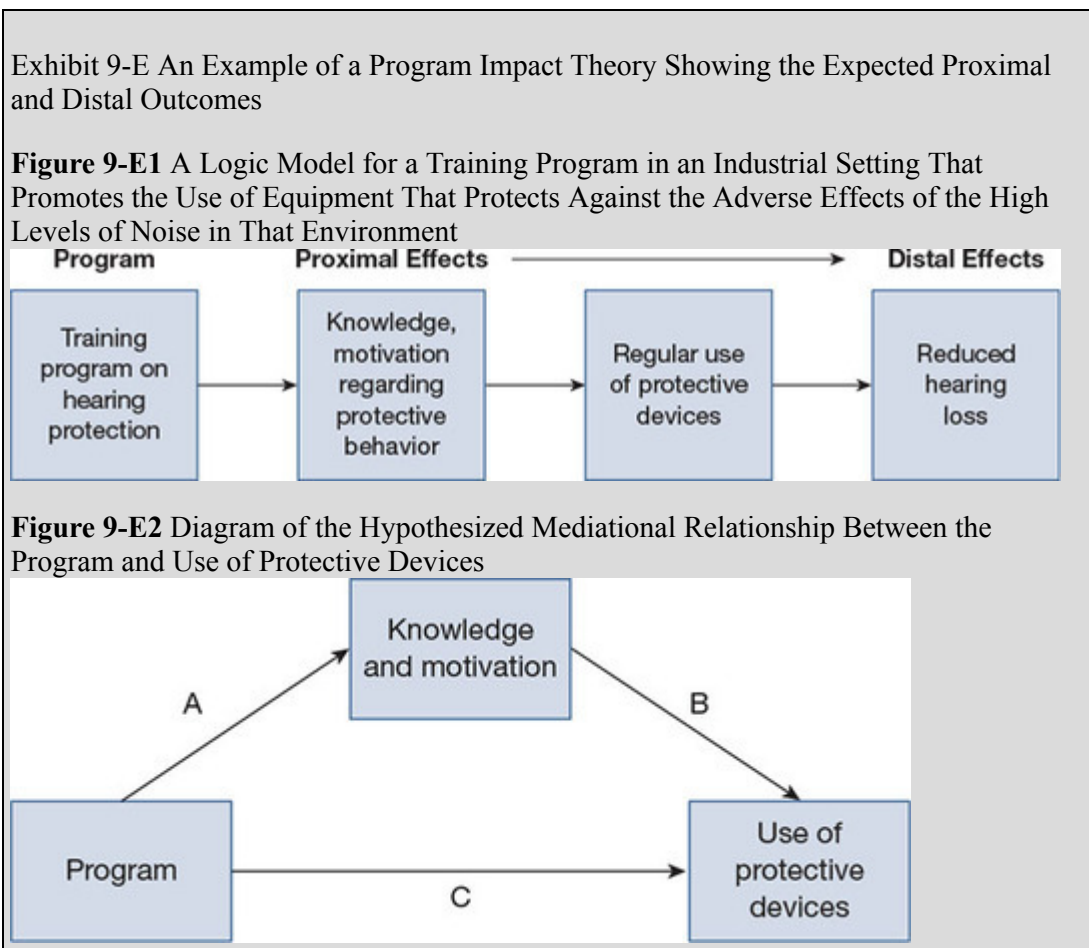
Mediator Analysis

Another aspect of variation in program effects that may warrant attention in an impact assessment concerns possible mediator relationships among outcome variables. A **mediator variable** in this context is a proximal outcome that changes as a result of exposure to the program and then, in turn, influences a more distal outcome. A mediator is thus an intervening variable that comes between program exposure and some key outcome with variation on that intervening variable correlated with variation on the key outcome. Mediator variables, therefore, represent a step along the causal pathway by which the program is expected to bring about change in the distal outcome. The proximal outcomes identified in a program's action theory, as discussed in [Chapter 3](#), are all conceptualized in those theories as mediator variables.

Like moderator variables, mediator variables are interesting for two reasons. First, exploration of mediator relationships helps the evaluator and the program stakeholders better understand what change processes occur among participants as a result of exposure to the program. This, in turn, can lead to informed discussion of ways to enhance that process and improve the program to attain better effects. Second, testing for the mediator relationships hypothesized in the program logic is another way of probing the evaluation findings to determine if they are fully consistent with what is expected if the program is in fact having the intended effects.

Exhibit 9-E illustrates the analysis of mediator relationships for a training program on the use of hearing protection devices in an industrial environment. The distal outcome the program intends to affect is job-related hearing loss. The causal pathway posited in the impact theory ([Figure 9-E1](#)) is that the training will produce increased knowledge about the adverse effects of environmental noise and motivation to use the protective equipment available in the workplace. That, in turn, is expected to lead to more actual use of the protective devices and then finally to reduced hearing loss. In this hypothesized pathway, knowledge and motivation are mediating variables between program exposure and use of the protective gear. Use of the protective gear, similarly, is presumed to mediate the

relationship between increased knowledge and motivation and reduced hearing loss.



To simplify, we will consider for the moment only the hypothesized role of knowledge and motivation as mediators of the effects of the training program on the use of the hearing protection devices. This relationship is shown in [Figure 9-E2](#), in which Path A-B-C represents the mediational relationship. A test of whether there are mediator relationships among these variables involves, first, confirming that there are program effects on both the proximal outcome (Path A) and the more distal outcome (Path C). If the proximal outcome is not influenced by the program, it cannot function as a mediator of the program influence on the more distal outcome. If the distal outcome does not show a program effect, there is nothing to mediate, but it can still be helpful to test the mediation because some mediators actually suppress, rather than enhance, the effects of a program. The critical test of mediation is whether the effects on the proximal outcome are related to the

effects on the distal outcome, in this example, whether variation in knowledge and motivation predicts variation in the use of the protective devices.

Detailed guidance on the statistical procedures for testing mediator relationships can be found in MacKinnon (2008) and VanderWeele (2015, 2016). For present purposes, our interest is mainly what can be learned from such analyses. If all the mediator relationships posited in the impact theory for the hearing protection training program were demonstrated by such analysis, this would add considerably to the plausibility that the program was indeed producing the intended effects. But even if the results of the mediator analysis were different from expectations, they will have diagnostic value for the program. Suppose, for instance, that employees' knowledge and motivation show an effect of the program but not a mediator relationship for using the protective gear. This pattern of results suggests that although knowledge and motivation are affected by the training program, they do not have much influence on the actual protective behavior. Thus, the program should be encouraged to explore more deeply the factors that are related to use of the protective devices. This might, for example, reveal that it has mostly to do with the extent to which the protective equipment interferes with performance of the tasks required of the employees. In that case, the program is likely to achieve better results if it makes appropriate changes in the protective gear and/or the way the respective tasks are to be performed.

The Role of Meta-Analysis

Thousands of impact evaluations of social, behavioral, economic, and public health programs have been conducted and reported in professional journals and sources available on the Internet. Familiarity with prior evaluation research in relevant program areas is important for impact evaluators. Prior evaluations can be instructive about successful evaluation designs in various circumstances, outcome measures responsive to program effects, the magnitude of effects that it is realistic to expect, the problems encountered conducting impact evaluations in pertinent program domains, and much more. Research reviews are common in many program areas and can provide informative overviews. For many of the evaluator's purposes, however, the most useful summary may come from a **meta-analysis** that statistically synthesizes the findings of scores or even hundreds of prior impact assessments. This form of research synthesis has become so common that there are few program areas in which a meta-analysis has not been conducted for whatever evaluation studies are available.

In a typical meta-analysis, reports of all available impact assessment studies that meet prespecified criteria are first collected. The focus may be on a particular type of program for a particular condition (e.g., psychotherapy for eating disorders), a broad program domain with multiple outcomes (e.g., support programs for the elderly), or a type of outcome irrespective of the type of program for which it is measured (e.g., adolescent bullying). Additionally, there are typically standards for eligible methods, the nature of the study samples, and selected other study characteristics such as geographical location, recency of the research, and so forth.

Once all the reports of eligible studies have been collected, the intervention effects on the outcomes of interest are encoded as effect sizes using an effect size statistic of the sort shown in [Exhibit 9-A](#). Descriptive information about the evaluation methods, program participants, nature of the intervention, and other such particulars is also recorded in a systematic form. All of these data are put in a database, and various statistical analyses are conducted on the overall mean effects for different outcomes, the variation in effects, and the factors associated with that variation

(Borenstein, Hedges, Higgins, & Rothstein, 2009; Lipsey & Wilson, 2001). The results can be informative for evaluators designing impact assessments of programs similar to those represented in the meta-analysis. In addition, by summarizing what evaluators collectively have found about the effects of various social interventions, the results can be informative to the field of evaluation. We turn now to a brief discussion of each of these contributions.

Informing an Impact Assessment

Any meta-analyses conducted and reported for interventions of the same general type as one for which an evaluator is planning an impact assessment will generally provide useful information for the design of that study. Consequently, the evaluator should pay particular attention to locating relevant meta-analysis work as part of the general review of the relevant literature that should precede an impact assessment. [Exhibit 9-F](#) summarizes results from a meta-analysis of school-based programs to prevent aggressive behavior that illustrate the kind of information often available.

Meta-analysis focuses mainly on the statistical effect sizes generated by intervention studies and thus can be particularly informative with regard to that aspect of an impact evaluation. When configuring an impact evaluation design for statistical power, for instance, an evaluator must have some idea of the magnitude of the effect size a program might produce and what MDES is worth trying to detect. Meta-analyses will typically provide information about the overall mean effect size for a program area and, often, breakdowns for different program variations. With information on the standard deviation of the effect sizes, the evaluator will also have some idea of the breadth of the effect size distribution and, hence, some estimate of the likely lower and upper range that might be expected from the program to be evaluated.

Program effect sizes, of course, may well be different for different outcomes. Many meta-analyses examine the different categories of outcome variables represented in the available evaluation studies. This information can give an evaluator an idea of what effects other studies have considered and what they found. Of course, the meta-analysis will be of less use if the program to be evaluated is concerned about an outcome that has not yet been examined in evaluations of other similar programs. Even then, however, results for similar types of variables—attitudes, behavior, achievement, and so forth—may help the evaluator anticipate both the likelihood of effects and the expected magnitude of those effects.

Similarly, after completing an impact evaluation, the evaluator may be able to use relevant meta-analysis results in appraising the magnitude of the program effects that have been found in the study. The effect size data presented by a thorough meta-analysis of impact evaluations in a program area constitute a set of norms that describe both typical program effects and the range over which they vary. An evaluator can use this information as a basis for judging whether the various effects discovered for the program being evaluated are representative of what similar programs attain. Of course, this judgment must take into consideration any differences in intervention characteristics, clientele, and circumstances between the program at hand and those represented in the meta-analysis results.

Exhibit 9-F An Example of Meta-Analysis Results: Effects of School-Based Intervention Programs on Aggressive Behavior

Many schools have programs aimed at preventing or reducing aggressive and disruptive behavior. To investigate the effects of these programs, a meta-analysis of the findings of 221 impact evaluation studies of such programs was conducted. A thorough search was made for published and unpublished study reports that involved school-based programs implemented in one or more grades from preschool through the last year of high school. To be eligible for inclusion in the meta-analysis, the study had to report outcome measures of aggressive behavior (e.g., fighting, bullying, person crimes, behavior problems, conduct disorder, acting out) and meet specified methodological standards. Standardized mean difference effect sizes were computed for the aggressive behavior outcomes of each study. The mean effect sizes for the most common types of programs were as follows:

<i>Therapy or counseling services</i> (therapy-like services such as group or individual counseling and case management)	.33
<i>Social competence training, cognitive-behavioral</i> (training to develop social skills, understand and control anger, resolve conflict, and the like using cognitive-behavioral approaches)	.27
<i>Behavioral and classroom management techniques</i> (use of various behavioral techniques, such as rewards, token economies, and contingency contracts, to shape behavior)	.22
<i>Social competence training, not cognitive behavioral</i> (training to develop social skills, understand and control anger, resolve conflict, and the like using instructional approaches other than cognitive-behavioral techniques)	.20
<i>Multimodal programs</i> (interventions that include at least three components such as social competence training, counseling, classroom management, parent training, academic services, and the like)	.15

In addition, a moderator analysis of the effect sizes showed that program effects were larger when

- high-risk children were the target population,
- programs were well implemented,
- programs were administered by teachers, and
- a one-on-one individualized program format was used.

Source: Adapted from Wilson, Lipsey, and Derzon (2003).

Furthermore, a meta-analysis that systematically explores the relationship between program characteristics and effects on different outcomes not only will make it easier for the evaluator to compare effects but may offer some clues about what features of the program are most critical to its effectiveness. The meta-analysis summarized in [Exhibit 9-F](#), for instance, found that programs were much less effective if they were delivered by laypersons (parents or volunteers) than by teachers and that better results were produced by a one-on-one than by a group format. An evaluator conducting an impact assessment of a school-based aggression prevention program might, therefore, want to pay particular attention to these characteristics of the program.

Informing the Evaluation Field

Aside from supporting the evaluation of specific programs, a major function of the evaluation field is to summarize what evaluations have found generally about the characteristics of effective programs. Though every program is unique in some ways, this does not mean that we should not aspire to discover some patterns in our evaluation findings that will broaden our understanding of what works, for whom, and under what circumstances. Reliable knowledge of this sort not only will help evaluators to better focus and design each program evaluation they conduct, but it will provide a basis for informing decision makers about the best approaches to ameliorating social problems.

Meta-analysis has become one of the principal means for synthesizing what evaluators and other researchers have found about the effects of social intervention in general. To be sure, generalization is difficult because of the complexity of social programs and the variability in the results they produce. Nonetheless, steady progress is being made in many program areas to identify more and less effective intervention models, the nature and magnitude of their effects on different outcomes, and the most critical determinants of their success. As a side benefit, much is also being learned about the role of the methods used for impact assessment in shaping the results obtained.

One important implication for evaluators of the ongoing efforts to synthesize impact evaluation results is the necessity to fully report each impact evaluation so that it will be available for inclusion in meta-analysis studies. In this regard, the evaluation field itself becomes a stakeholder in every evaluation. Like all stakeholders, it has distinctive information needs that the evaluator must take into consideration when designing and reporting an evaluation.

Summary

- The ability of an impact assessment to detect program effects, and the importance of those effects, depends in large part on their magnitude. An impact evaluation estimates statistical effects on the target outcomes that can be described in various

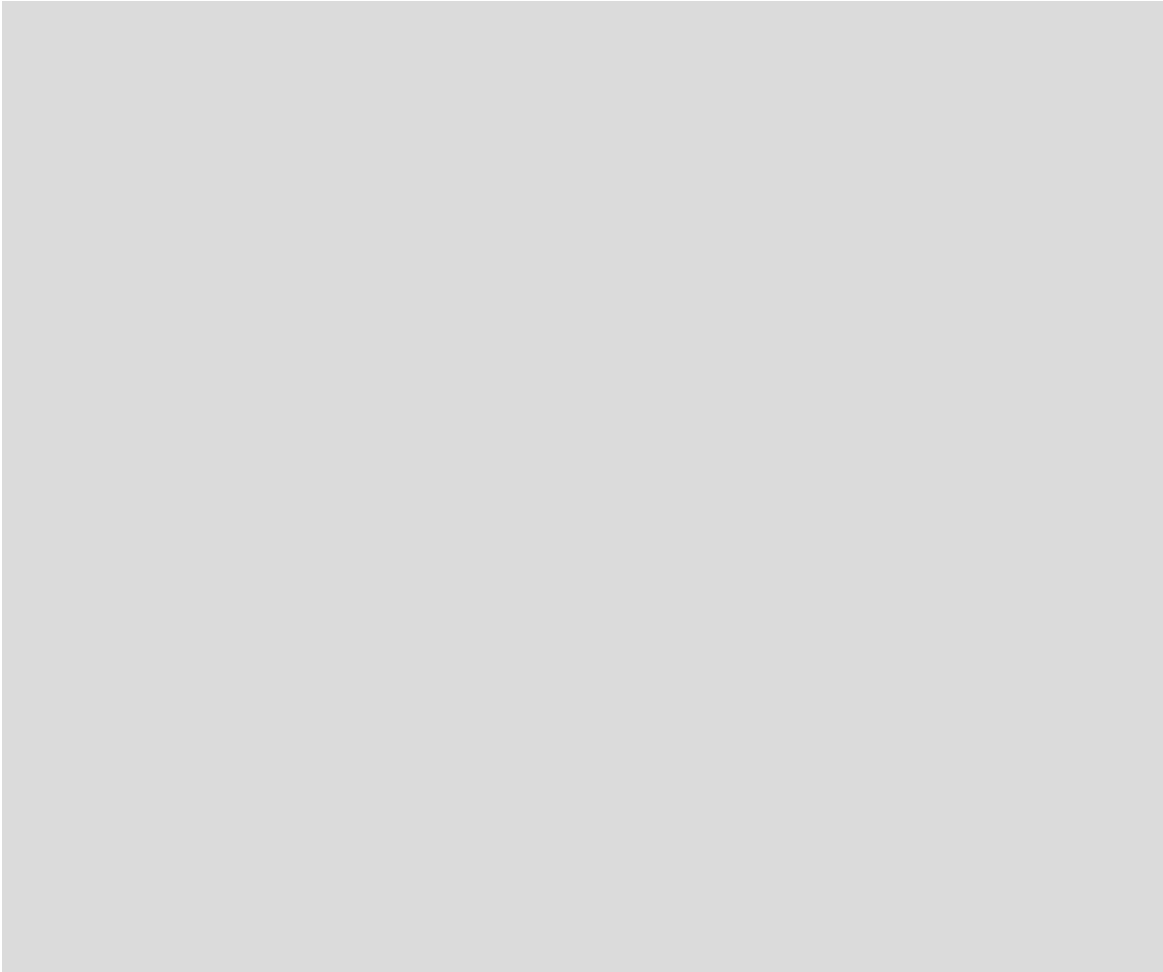
ways, including with standardized effect size statistics that allow comparisons across outcomes and studies.

- The most commonly used standardized effect size statistics are the standardized mean difference for continuous outcome measures and the odds ratio for binary outcome measures.
- Impact evaluations produce statistical effect size estimates that are not necessarily the true effect sizes because of various chance factors that contribute statistical noise to the estimates. What it means to detect a program effect under these circumstances is that an appropriate statistical test indicates that the observed effect size is statistically significant, that is, it is unlikely to have occurred simply by chance.
- It can be difficult to detect small program effects at a statistically significant level, and effects so small that they do not represent meaningful change in the relevant outcomes have little practical value. A critical step in the design of an impact evaluation, therefore, is specifying the smallest effect size that has practical significance in the context of the program and its target outcomes. This is referred to as the minimum detectable effect size (MDES).
- There is no single best way to identify an MDES that is at the threshold of practical significance. It cannot be done simply on the basis of the numerical value of the effect size statistic; it requires a translation of the effect size on a given outcome into terms that allow interpretation of its practical implications.
- An impact evaluation should be designed to have a high probability of finding statistical significance for program effects if they are as large as or larger than the MDES. The statistical framework for developing that design revolves around the concept of statistical power, which is defined directly as the probability of statistical significance when there is a true effect of a given magnitude.
- The four factors that determine statistical power are: (a) the effect size to be detected (the MDES), (b) the alpha level for statistical significance (conventionally .05), (c) the sample size, and (d) the statistical significance test used. Sample size is the major factor over which the evaluator has influence, but the sample size required when the MDES is modest can be very large. Including baseline covariates highly correlated with the outcome in the statistical significance test can appreciably reduce the sample size needed and is a useful technique.
- Whatever the overall mean program effect, there are usually variations in effects for different subgroups of the target population. Investigating moderator variables that identify distinct subgroups is an important aspect of impact assessment. This may reveal that program effects are especially large or small for some subgroups, and it allows the evaluator to probe the outcome data in ways that can strengthen the overall conclusions about a program's effectiveness.
- The investigation of mediator variables probes variation in proximal program effects in relationship to variation in more distal effects to determine if one leads to the other as implied by the program's impact theory. These linkages define mediator relationships that can inform the evaluator and stakeholders about the change processes that occur among targets as a result of exposure to the program.
- The results of meta-analyses can be informative for evaluators designing impact assessments. Their findings typically identify the outcomes affected by the type of program represented and the magnitude and range of effects that might be expected on those outcomes. This information can help identify relevant outcomes and provide a realistic perspective about the effects likely to occur and plausible MDES values.

- In addition, meta-analysis has become the principal means for synthesizing what evaluators and other researchers have found about the effects of social intervention. In this role, it informs the evaluation field about what has been learned collectively from the thousands of impact evaluations that have been conducted over the years.

Key Concepts

Effect size statistic 213
Effective sample size 224
Mediator variable 229
Meta-analysis 231
Minimum detectable effect size (MDES) 216
Moderator variable 226
Odds ratio 213
Sampling error 220
Standardized mean difference 213
Statistical power 221
Type I error 222
Type II error 221



Critical Thinking/Discussion Questions

1. Describe the two most commonly used standardized effect size statistics and explain when each one is appropriate to use.
2. Define the minimum detectable effect size (MDES) and explain how to determine an appropriate MDES.
3. Explain what a mediator variable is and how it can affect more distal outcomes. Provide an example of a mediating variable in a relationship between program exposure and a specific outcome.

Application Exercises

1. Locate a thorough evaluation report that measures program effects. Discuss what statistical tests were used to calculate program effects. Specify the valence and magnitude of the statistical findings in a sentence describing the program effects on one outcome variable.
2. Find a meta-analysis of impact assessments for an intervention domain. Produce a short summary of the meta-analysis focusing on the criteria for inclusion of impact evaluations in the analysis and the findings of the meta-analysis.

Chapter 10 Assessing the Economic Efficiency of Programs

Key Concepts in Efficiency Analysis

Ex Ante and Ex Post Efficiency Analyses

Cost-Benefit and Cost-Effectiveness Analyses

Conducting Cost-Benefit Analyses

Assembling Cost Data

Accounting Perspectives

Measuring Costs and Benefits

Monetizing Benefits

Estimating Costs

Other Considerations in Cost-Benefit Analysis

Comparing Costs With Benefits

When to Do Ex Post Cost-Benefit Analysis

Conducting Cost-Effectiveness Analyses

Summary

Key Concepts

Whether programs have been implemented successfully and the degree to which they are effective are at the heart of evaluation. However, it is also important for stakeholders to be informed about the cost required to obtain a program's effects and whether those benefits justify the costs. Comparison of the costs and benefits of social programs is one of the most relevant considerations in decisions about whether to continue, expand, revise, or terminate them.

Efficiency assessments—cost-benefit and cost-effectiveness analyses—provide a frame of reference for relating costs to program impacts. In addition to providing information for making decisions on the allocation of resources, they are often useful in gaining the support of planning groups and political constituencies that determine the fate of social intervention efforts.

The procedures used in both types of analyses can be quite technical, and this chapter provides only a broad overview of their application illustrated with examples. However, because the issue of the cost required to achieve a given magnitude of desired change is implicit in all impact evaluations, it is important for evaluators to understand the ideas embodied in efficiency analyses and their relevance to the task of fully accounting for a program's social value.

Efficiency issues arise frequently in decision making about social interventions, as the following examples illustrate.

- Policymakers must decide whether to allocate funding to a basic literacy program for new immigrants that has shown positive effects in an impact evaluation. An important consideration is the extent to which the program's **benefits** (positive outcomes, both direct and indirect) exceed its **costs** (direct and indirect inputs required to produce the intervention).
- A government agency is reviewing national disease control programs currently in operation. If additional funds are to be allocated to disease control, the administrators want to know which programs would show the biggest payoff per dollar of expenditure.
- Evaluations in criminal justice have established the effectiveness of various alternative programs for reducing recidivism. The most effective program is also the most costly. The question for the decision makers is whether the greater effectiveness of that program justifies its higher cost.
- Board members of a private foundation are debating whether to support a program of low-interest loans for home purchases or a program to provide work skills training for married women to increase family income. They want to know which will produce the greatest economic benefit for low-income families.

These are examples of the kind of resource allocation issues commonly faced by planners, funders, and policymakers. Again and again, decision makers must choose how to allocate limited funds to put them to optimal use. Consider even the fortunate case in which pilot projects of several programs have shown them all to be effective in producing the desired effects. The decision of which to fund on a larger scale must take into account the relationship between those effects and the cost of producing them. Although other factors, including political considerations, come into play, the preferred program often is the one that produces the most impact for a given level of expenditure. This simple principle is the foundation of cost-benefit and cost-effectiveness analyses, techniques that provide systematic approaches to the analysis of resource allocations.

Both cost-benefit and cost-effectiveness analyses are means of judging the economic efficiency of programs. As we will elaborate, the difference between these two types of analysis is the way the effects of a program are expressed. In *cost-benefit* analyses, the outcomes affected are expressed in monetary terms; in *cost-effectiveness* analyses, the outcomes affected are expressed in substantive terms. For example, a cost-benefit analysis of a program to reduce cigarette smoking might focus on the difference between the dollars expended on the antismoking program and the dollar savings from reduced medical care for smoking-related diseases. A cost-effectiveness analysis of the same program might estimate the cost associated with the conversion of one smoker into a nonsmoker.

The idea of judging the utility of social intervention efforts in terms of their economic efficiency has gained widespread acceptance. However, the question of “correct” procedures for conducting cost-benefit and cost-effectiveness analyses of such programs remains an area of controversy. This controversy is related to a combination of the need for judgment calls about the data and analytical procedures used, reluctance to impose monetary values on many social program effects, the uncertainty of how to weigh current costs against future benefits, and an unwillingness to forsake initiatives that have been held in esteem for extended periods of time despite their cost. Evaluators undertaking cost-benefit or cost-effectiveness analyses of social interventions must be aware of the particular issues involved in applying efficiency analyses, as well as the limitations that characterize the use of cost-benefit and cost-effectiveness analyses in general. (For comprehensive discussions of economic efficiency assessment procedures, see Boardman, Greenberg, Vining, & Weimer, 2018; Levin & McEwan, 2001; Mishan & Quah, 2007.)

Key Concepts in Efficiency Analysis

Cost-benefit and cost-effectiveness analyses can be viewed both as conceptual perspectives and as technical procedures. From a conceptual point of view, efficiency analysis asks that we think in a disciplined fashion about both costs and benefits. In the case of virtually all social programs, identifying and comparing the actual or anticipated costs with the known or expected benefits can prove invaluable. Most other types of evaluation focus mainly on the benefits. Furthermore, efficiency analyses provide a comparative perspective on the relative utility of interventions. Judgments of comparative utility are unavoidable given that most social programs are conducted under resource constraints. A salient illustration of a contribution to decision making along these lines is a cost-effectiveness analysis of two interventions for reducing the incidence of HIV/AIDS infections among Kenyan teenagers (see [Exhibit 10-A](#)). As the report of this analysis documents, both interventions were effective, but one was much more cost-effective than the other.

Despite their potential value, we want to emphasize that the results from cost-benefit and cost-effectiveness analyses should be viewed with caution, and sometimes with a fair degree of skepticism. Expressing the results of an evaluation study in efficiency terms may require taking into account different costs and outcomes depending on the perspectives and values of sponsors, stakeholders, and beneficiaries. And cost estimates can be made in different ways, what are referred to as **accounting perspectives** (discussed later in this chapter). Furthermore, efficiency analysis is often dependent on at least some untested assumptions, and the requisite data may not be fully available. In some applications, the results may show unacceptable levels of sensitivity to reasonable variations in the analytic and conceptual models used and their underlying assumptions. These features can make the results of the most careful efficiency analysis arguable or even unacceptable to some stakeholders who disagree with the perspective taken by the analyst. Even the strongest advocates of efficiency analyses rarely argue that such studies should be the sole determinant of decisions about programs. Nonetheless, they are a valuable input into the complex mosaic from which decisions emerge.

Ex Ante and Ex Post Efficiency Analyses

Efficiency analyses are most commonly undertaken either prospectively during the planning and design phase for a program (**ex ante efficiency analysis**) or retrospectively, after a program is in place and has been demonstrated to be effective by an impact evaluation (**ex post efficiency analysis**). In the planning and design phases, ex ante efficiency analyses are undertaken on the basis of a program's anticipated costs and outcomes. Such analyses, of course, must assume a given magnitude of positive impact even if it is based only on conjecture. Likewise, the costs of providing and delivering the intervention must be estimated. Because ex ante analyses cannot be based entirely on actual program implementation costs and effects, they risk under- or overestimating the program's economic efficiency.

Exhibit 10-A Cost-Effectiveness of Two Educational Interventions for Reducing HIV/AIDS

Sub-Saharan Africa has the highest rate of HIV infection in the world. About one fourth of infections occur in people under the age of 25, nearly all as a result of unprotected sex with teenage girls, among the most vulnerable. Randomized impact evaluations conducted in Kenya have demonstrated the effectiveness of two programs for reducing the incidence of unprotected sex among teenagers, and of pregnancy among teenage girls: the Relative Risk program and the Uniform Subsidy program.

The Relative Risk program provides eighth grade students with detailed HIV risk information through presentations made during visits to their schools by trained project officers that include a video and time for group discussion. The emphasis in this educational intervention is on intergenerational sex: men over the age of 25 and teenage girls. The Uniform Subsidy program provides two free school uniforms to students in each of the last 3 years of primary school (Grades 6–8), during which dropout rates are especially high. The free uniforms reduce the cost of school attendance, with the objective of keeping students in school longer and offsetting the higher risk for pregnancy among girls who drop out. The impact evaluations found that the Relative Risk program reduced the incidence of childbearing by 1.5%, and the Uniform Subsidy program reduced the childbearing rate by 2.7%, assessed at 1 year after the end of each program.

The cost-effectiveness analysis first identified the inputs required to operate each program through a review of program documents, discussions with program personnel, and direct observations of the interventions. The cost of each such item was then estimated using local retail prices, salaries for personnel prorated for time invested in the respective program, and the school support cost for the required classroom time (with inflation adjustments for the cost estimates that were not contemporaneous).

The total cost of the Relative Risk program was 161,151 Kenyan shillings (KES) for 1,212 participating girls, yielding a cost per student of KES 133. The total cost of the Uniform Subsidy program was KES 2,603,753 for 1,250 participating girls, for a cost per student of KES 2,083. The most relevant comparison, however, was on the cost per pregnancy averted. For the Relative Risk program, the impact estimate was 18 pregnancies averted at a cost of KES 8,864 each. The Uniform Subsidy program impact estimate was 34 pregnancies averted at a cost of KES 77,148 each. Although the Uniform Subsidy program was more effective in reducing teen pregnancies, the cost per pregnancy averted for the Relative Risk program was far less, making it the more cost-effective program.

Source: Mustafa (2018).

Ex ante cost-benefit analyses are most important for programs that will be difficult to abandon once they have been put into place, or that require extensive commitments and resources to be realized. For example, the decision to increase recreational beach facilities by putting in new jetties along the New Jersey coastline would be difficult to overturn once the jetties had been constructed. Therefore, there is a need to develop reasonable estimates of the costs and benefits of such an initiative compared with other ways of increasing recreational opportunities.

Exhibit 10-B illustrates an ex ante analysis to assess whether it would likely be cost-beneficial for health insurers to reimburse patients for home-based blood pressure monitoring. Home monitoring has been shown to be more effective than clinic monitoring, but it does not follow that the cost to insurers of reimbursement would result in sufficient savings in other medical expenditures to justify provision of insurance coverage for this procedure. Unusually good data were available for this cost-benefit analysis from prior evaluations of home monitoring versus clinic monitoring and extensive medical records relating blood pressure diagnostic information to medical treatment.

Exhibit 10-B Ex Ante Cost-Benefit Analysis From the Perspective of Insurers for Home Blood Pressure Monitoring

Hypertension is a significant risk factor for cardiovascular diseases and a primary contributor to health care expenditures, and accurate blood pressure measurement is essential for its diagnosis and treatment. Blood pressure monitoring during clinic visits is the most common method for diagnosing hypertension but is subject to error because of atypical readings resulting from patients' reactions to that medical environment. An alternative is self-monitoring by the patient at home, which has been shown to be more effective than clinical monitoring for diagnosing and managing hypertension.

Despite its effectiveness, most U.S. insurers do not reimburse for the equipment and training required for home monitoring under the belief that it is not cost-beneficial from the insurer's perspective. It requires up-front costs, with the marginal benefits it has beyond standard health care not likely to appear for many years.

The cost-benefit analysis summarized here was based on the insurance claims records for 16,375 members of two health insurance plans (a private employee plan and a Medicare Advantage plan) with a diagnosis of hypertension. The claims data were used to estimate the transition probabilities from an initial physician visit to hypertension diagnosis, to treatment, to hypertension-related cardiovascular diseases, and finally to patient death and the costs to the insurer at each of these stages. Clinic-based blood pressure monitoring was the standard of care in these data. To estimate the transition probabilities with home monitoring, the clinic monitoring probabilities were adjusted for the effectiveness of home monitoring relative to clinic monitoring reported in a meta-analysis based primarily on randomized prospective studies making this comparison.

Reimbursement costs to the insurer for home monitoring were assumed to include the cost of the blood pressure monitoring devices plus the costs of an awareness-raising campaign to educate members of the health plans and their physicians about their availability. The equipment costs were based on retail prices discounted for wholesale purchase with an assumed lifetime of 5 years. All costs and benefits were expressed in current dollars, taking into account the diminishing value of dollars spent or saved in the future.

For the employee health plan, home monitoring was estimated to yield overall net savings beyond the cost of reimbursement in the 1st year of \$33.75 per member aged 20 to 44 years and \$32.65 per member aged 45 to 64 years. By year 10 these net savings had increased to \$414.81 per member aged 20 to 44 years and \$439.14 per member aged 45 to 64 years. For members of the Medicare Advantage plan aged ≥ 65 years, 1st-year net savings were \$166.17 per member and increased to \$1,364.27 per member by year 10.

These findings indicate that reimbursement of home blood pressure monitoring by an insurance company would be expected to generate overall net savings for the company as early as the 1st year and that these savings would grow larger over time.

Source: Adapted from Arrieta, Woods, Qiao, and Jay (2014).

Most commonly, efficiency analyses of social programs take place after completion of an impact evaluation. In such ex post cost-benefit and cost-effectiveness analyses, the objective is usually to determine whether the magnitude of the program effects are sufficient to justify the costs of the intervention. The focus of such assessments may be on the efficiency of a program in absolute or comparative terms, or both. In absolute terms, the idea is to judge whether the program is worth what it costs either by comparing costs to the monetary value of benefits or by calculating the money expended to produce some unit of outcome of recognized value. For example, a cost-benefit analysis may reveal that for each dollar spent to

reduce shoplifting in a department store, \$2 are saved in terms of stolen goods. Alternatively, a cost-effectiveness study might show that the program expends \$50 to avert each shoplifting incident.

In a comparative analysis, the issue is to determine the differential payoff of one program versus another, for example, comparing the costs of elevating the reading achievement scores of schoolchildren by one grade level produced by a computerized instruction program with the costs of achieving the same increase through a peer tutorial program. [Exhibit 10-A](#) also presents an example of such a comparative analysis.

Cost-Benefit and Cost-Effectiveness Analyses

Many considerations besides economic efficiency are brought to bear in policy making, but economic efficiency is almost always relevant. Cost-benefit and cost-effectiveness analyses have the virtue of encouraging evaluators to become knowledgeable about program costs. Surprisingly, evaluators often pay little attention to those costs even though cost is a very salient issue for many stakeholders with influence over decisions about the program.

A cost-benefit analysis requires estimates of the benefits of a program and estimates of the costs of undertaking the program. Once specified, the benefits and costs are translated into common monetary units so they can be compared. Any cost-benefit analysis requires that a particular economic perspective be adopted and that certain assumptions be made to translate program inputs and outputs into monetary values. The economic perspective taken and the assumptions made will influence the resulting conclusions. Consequently, the analyst should, at the very least, state the basis for the assumptions that underlie the analysis. Often, analysts do more than that. For example, they may undertake sensitivity analyses that alter important assumptions to test how sensitive the findings are to variations in those assumptions—a central feature of a well-conducted efficiency study.

For social programs, there is generally more concern about converting outcomes into monetary values than there is about inputs. Cost-benefit analysis is least controversial when applied to technical and industrial projects for which it is relatively easy to place a monetary value on benefits. Examples include engineering projects designed to reduce the costs of electricity to consumers, highway construction to facilitate transportation of goods, and irrigation improvements to increase crop yields. Estimating the benefits of social programs in monetary terms is frequently difficult because those benefits often have a social value not easily captured in economic terms. For example, future occupational benefits from vocational training may be translated into the monetary value of increased earnings in a relatively straightforward and uncontroversial manner. The issues are more problematic with such social interventions as fertility control programs or

health services because one must ultimately place a value on human life to fully monetize the program benefits (Nyborg, 2014). Even short of issues of life and death, it is difficult to place a monetary value on such outcomes as learning to read, improving marital relationships, overcoming depression, and raising healthy children.

Because of the controversial nature of valuing outcomes, cost-effectiveness analysis is often seen as a more appropriate technique than cost-benefit analysis for an efficiency analysis of many social programs. Cost-effectiveness analysis requires monetizing only the program's costs; its benefits are expressed in outcome units. For example, the cost-effectiveness of distributing free textbooks to rural primary school children could be expressed in terms of how much the average reading scores of the students increased for each \$1,000 in program costs. For cost-effectiveness analysis, then, efficiency is expressed in terms of the costs of achieving a given result. [Exhibit 10-C](#) describes such a case. In that example, a cost-effectiveness analysis of a weight-loss program measured benefits in terms of the gain in quality-adjusted life-years per person attained over different postprogram periods for a program cost of \$846 per person.

Exhibit 10-C Cost-Effectiveness of a Weight Control Intervention Designed for Mexican Americans

Ethnic minorities in the United States are disproportionately affected by obesity and diabetes. For example, among Mexican Americans, 74% of men and 72% of women are overweight, and their rates of Type 2 diabetes are twice those of non-Hispanic Whites. A total of 519 men and women from a Mexican-origin population residing along the Texas-Mexico border participated in Beyond Sabor, a 12-week, culturally tailored, community-based weight-control program designed to reduce risk factors for obesity and diabetes. An impact evaluation found that 34% of those who completed the program achieved 2% weight loss, and 14% achieved 5% weight loss.

For the cost-effectiveness analysis, program costs were calculated to include all input to the program, including time and transportation costs for the participants as well as staff and supply costs for program delivery. That estimate was a total program cost of \$846 per person. Program outcomes were represented in terms of the quality-adjusted life-years (QALYs) saved by the intervention. QALYs are a measure of the value of health outcomes often used in medical contexts. They combine length of life and quality of life into a single index number. One QALY represents 1 year in perfect health; with poorer health, the figure is adjusted downward, reaching zero for death.

A validated software program was used to project the program's lifetime health outcomes on the basis of the proportions of participants achieving the 2% and 5% weight-loss goals.

The table below presents the QALYs per person gained on average at an average cost of \$846 per person over different postintervention periods for participants meeting each of those goals.

Quality-Adjusted Life-Year Gains Per Person

Postintervention Period	2% Weight Loss	5% Weight Loss
5 years	.14	.30
10 years	.35	.81
20 years	.91	2.11

Source: Adapted from Wilson, Brown, and Bastida (2015).

Source: Adapted from Wilson, Brown, and Bastida (2015).

Conducting Cost-Benefit Analyses

With the basic concepts of efficiency analysis in mind, we turn to how such analyses are conducted. Because many of the basic procedures are similar, we discuss cost-benefit analysis in detail and then treat cost-effectiveness analysis more briefly. We begin with a step necessary to both types of studies, assembling the cost data.

Assembling Cost Data

Cost data are obviously essential to economic efficiency analyses. In the case of ex ante analyses, program costs must be estimated on the basis of costs incurred in similar programs or on knowledge of the costs of the relevant program components. For ex post efficiency analyses, it is necessary to analyze program financial data, segregating out the funds used to finance program processes as well as collecting costs incurred by participants or other associated agencies. Useful sources of cost data include the following:

- Agency fiscal records: These include salaries of program personnel, space rental, stipends paid to clients, supplies, maintenance costs, business services, and so on.
- Participant cost estimates: These include imputed costs of time spent by clients in program activities, client transportation costs, and so on. (Typically these costs must be estimated.)
- Cooperating agencies: If a program includes activities of a cooperating agency, such as a school, a health clinic, or another government agency, the costs borne may be obtained from the cooperating agency.

Fiscal records, it should be noted, are not always easily interpreted for purposes of efficiency analysis. The evaluator may have to seek help from an accounting or a financial professional. It is often useful to draw up a list of the cost data needed for a program. [Exhibit 10-D](#) shows a worksheet representing the various costs for a program that provided high school students with exposure to working academic scientists to heighten students' interest in pursuing scientific careers. Note that the worksheet identifies the several parties to the program who bear program costs.

Accounting Perspectives

To carry out a cost-benefit analysis, one must first decide what perspective to take in calculating costs and benefits, that is, the point of view that should be the basis for specifying, measuring, and monetizing benefits and costs and determining which costs and benefits are included. Benefits and costs must be defined from a single perspective because mixing points of view would result in confused specifications and overlapping or double counting. Of course, more than one cost-benefit analysis for a single program may be undertaken, each from a different perspective. Separate analyses based on different perspectives often provide information on how benefits compare with costs as they affect different relevant stakeholders. Generally, one or more of three accounting perspectives are used for analysis of social programs, those of (a) individual participants or target populations, (b) program sponsors, and (c) the communal social unit that provides the context and, perhaps, some support for the program (e.g., municipality, county, state, or nation).

Exhibit 10-D Worksheet for Estimating Annualized Costs for a Hypothetical Program

Saturday Science Scholars is a program in which a group of high school students gather for two Saturdays a month during the school year to meet with high school science teachers and professors from a local university. Its purpose is to stimulate the students' interest in scientific careers and expose them to cutting-edge research. The following worksheet shows the various program ingredients, their cost, and whether they were borne by the government, the university, or participating students and their parents.

Cost Ingredients	Total Cost	Cost to School District	Cost to State Government	Cost to University	Cost to Students and Parents
Personnel					
2 high school teachers	9,000	9,000			
2 university professors	14,400			14,400	
2 parent aides (volunteers)	3,600				3,600
Facilities					
High school science lab and classroom	2,000	2,000			
Materials and equipment					
Photocopies	400	400			
Materials for science experiments	500	250		250	
Laboratory equipment	500			500	
Other					
Maintenance and janitorial services	1,500	1,500			
Insurance	1,800	1,800			
Utilities	900	900			
Required participant inputs					
Transportation (time, vehicle costs)	625				625
Total ingredients cost	35,255	15,850	0	15,150	4,225
User fees		-1,000			1,000
Other cash subsidies		-7,500	7,500		
Net costs	35,225	7,350	7,500	15,150	5,225

Source: Adapted from Levin and McEwan (2001).
Source: Adapted from Levin and McEwan (2001).

The *individual-target population* accounting perspective takes the point of view of the persons, groups, or organizations receiving the intervention or program services. Cost-benefit analyses from this perspective often produce higher benefit-to-cost results (net benefits) than analyses using other perspectives. In particular, if the program sponsor or other social agents bear most of the cost and the program participants receive most of the benefits, the benefit-cost relationship for participants will be especially favorable. For example, an adult education program may impose relatively few costs on participants—primarily the time spent participating in the program. Furthermore, if the time required is mainly in the evenings, there may be no loss of income involved. The benefits to the participants, meanwhile, may include improvements in earnings, greater job satisfaction, and more occupational options. [Exhibit 10-E](#) describes a cost-benefit

analysis of an employment training program that shows much the same pattern.

The *program sponsor* accounting perspective takes the point of view of the funding source in valuing benefits and specifying cost factors. The funding source may be a private agency or foundation, a government agency, or a for-profit organization. From this perspective, the cost-benefit analysis is designed to reveal what the sponsor pays to provide a program and what benefits or savings accrue to that sponsor. The program sponsor accounting perspective is most appropriate when the sponsor must make choices between alternative programs. A county government, for example, may favor a vocational education initiative that includes student stipends over other programs because it reduces the costs of public assistance to the eligible unemployed participants. Also, if the future incomes of the participants increase because of the training, their increased tax payments would be a benefit to the county government. On the cost side, the county government incurs the costs of program personnel, supplies, facilities, and the stipends paid to the participants during the training.

Exhibit 10-F provides another example of a cost-benefit analysis conducted from the perspective of the program sponsors. It summarizes a study of the savings to the Medicaid and state behavioral health systems that result from implementation of community-based wraparound services for youth with serious emotional disturbances after their release from institutional care. Although implementation of that program involves new costs, the question of interest to the sponsoring health insurers is whether it lowers the cost of the claims they must pay for the health services used by the participating youth after they leave institutional care.

The *communal* or *social* accounting perspective takes the point of view of the community or society as a whole, usually in terms of total income. It is, therefore, the most comprehensive perspective but also usually the most complex and thus the most difficult to apply. Taking the point of view of society as a whole may require special efforts to account for **secondary effects**, or *externalities*—indirect project effects, whether beneficial or detrimental, on groups not directly involved with the intervention. A secondary effect of a training program, for example, might be the spillover

of the training to relatives, neighbors, and friends of the participants. Among the more commonly discussed negative external effects of industrial and technical projects are pollution, noise, traffic, and destruction of plant and animal life. Moreover, in the current literature, communal cost-benefit analysis has been expanded to include equity considerations, that is, the **distributional effects** of programs among different subgroups. Such effects result in a redistribution of resources in the general population. From a communal standpoint, for example, every dollar earned by a minority member who had been unemployed for 6 months or more may be seen as a “double benefit” and so entered into the analyses.

Exhibit 10-E Cost-Benefit Analyses of an Employment Training Program From the Participant and the Social Perspective

Accelerating Opportunity (AO) is a program aimed at helping adults with low basic skills earn industry-recognized credentials in high-growth occupations by allowing them to enroll in specially designed career and technical education courses at 2-year colleges without the usual prerequisites. Supportive services and connections with employers and workforce agencies facilitate completion of the coursework and transition to the workforce.

The evaluation team conducted quasi-experimental impact evaluations on the AO program in four states that estimated the earnings of participants over the 3 years after enrollment compared with those of a matched comparison group of nonparticipants. The comparison groups were matched with propensity score techniques within each state from adult education students who tested within the skill levels required for AO eligibility and had demographic, educational, and employment characteristics similar to those of the AO participants. Across the four states, 30 colleges and 4,572 students contributed data to the impact evaluation and cost-benefit analyses.

The cost-benefit analyses were conducted from two different perspectives. The individual participant perspective considered only costs incurred by the students and the benefits they received. The social perspective incorporated costs and benefits associated with all the actors involved in the program including, for instance, the colleges that hosted the AO training. In particular, student costs included their actual expenditures (e.g., tuition) as well as any forgone earnings while they were in school. Student benefits were their earnings gains relative to nonparticipants after taxes and reductions in social assistance. The social costs included the student costs plus the resource expenditures of the colleges for supporting AO (e.g., personnel) and the state administrative and oversight costs. The social benefits consisted of the total student earnings gains, assumed to represent increased productivity. From both perspectives, net benefits over the 3 years after AO enrollment were calculated by subtracting the costs from the benefits. The table below shows the net student and social benefit estimates over the 3 years for each state.

Net Benefits per Student for Each State

State	Net Student Benefit	Net Social Benefit
Illinois	705	-7,214
Kansas	4,030	1,595
Kentucky	-305	-3,922
Louisiana	1,639	-1,025

Note: Net benefits in 2015 dollars.

Note: Net benefits in 2015 dollars.

These results show that there was great variation across the states, but with net student benefits always larger than net social benefits, although negative for Kentucky. The net social benefits are negative for every state except Kansas, which incurred a relatively low cost per student and achieved a higher per-student benefit. The overall picture is that the colleges and state administration absorbed most of the cost of the AO program while the participating students reaped most of the benefits.

The authors describe some of the differences across the states that might account for the differences in net benefits. For example, Kansas had a particularly strong labor market for low-skill workers. And some states, such as Louisiana, had other employment training initiatives in the community college system that benefited the students in the comparison group more than in other states.

Source: Adapted from Kuehn et al. (2017).

Exhibit 10-F Costs and Savings to the Mental Health System of Providing Wraparound Services for Youth with Serious Emotional Disturbances

Treating youth with serious emotional disturbances (SEDs) often requires expensive institutional care. High Fidelity Wraparound (Wrap) is a support program designed to help sustain community-based placements for youth with SEDs through intensive, customized care coordination among parents, child-serving agencies, and providers. A number of controlled studies have demonstrated positive effects of Wrap on such outcomes as residential placements, mental health symptoms, school success, and juvenile justice recidivism.

This cost-benefit study was conducted in a southeastern state in the United States to assess the extent to which Wrap might reduce Medicaid and state behavioral health expenditures over a relatively long-term follow-up period after youth with SEDs were released from institutional care. A total of 161 youth transitioning from institutional care into Wrap were compared with a group of 324 youth who did not participate in Wrap after release from institutional care. Youth in both groups had a diagnosis that classified their mental illness as a serious emotional disturbance. The two groups were matched on the start date of their stay in institutional care and had similar functional assessment scores.

Total health care spending was determined from Medicaid and State Behavioral Health Authority claims data for the 12 months before Wrap participation and the combined time during and 12 months after participation in Wrap (average of 21 months), and for matching before and after periods for the youth in the comparison condition.

The youth who participated in Wrap were found, on average, to be younger than the youth in the control group, less likely to be in foster care, and to have required more health care spending per month during the 12-months before Wrap participation. To estimate Wrap effects on health expenditures during the follow-up period, a difference-in-differences regression analysis comparing pre-post expenditure differences for the Wrap and control group was conducted using the available baseline covariates and youth fixed effects.

The cost of the Wrap program for the participating youth averaged \$693/month over the follow-up period. The results of the regression analysis showed that Wrap participation was associated with total health expenditures that were \$1,823/month lower than those of control youth. This reduction stemmed largely from less use of mental health inpatient services during the follow-up period, as shown in the table below.

Estimates of the Differences in Monthly Health Expenditures for Wrap Participants Relative to Control Youth During the Follow-Up Period

Expenditure Category	Monthly Average Cost
Medical inpatient services	59
Medical outpatient services	-117
Mental health inpatient services	-2,137
Mental health outpatient services	331
Prescriptions	41
Total	-1,823

Over the average 21-month follow-up period, therefore, the cost savings associated with Wrap were approximately \$38,283 ($21 \times 1,823$), making Wrap quite cost-effective as a transition service for youth with serious mental disturbances released from institutional care.

Source: Adapted from Snyder, Marton, McLaren, Feng, and Zhou (2017).

The cost-benefit analyses on the Accelerating Opportunity employment training program summarized in [Exhibit 10-E](#) included an analysis developed from the social accounting perspective that contrasts with the one from the perspective of the program participants. From the social perspective, costs and benefits to the colleges that host the training program and the state agencies that administer and oversee it are taken into account as well as those for the program participant. [Exhibit 10-G](#) provides a simplified, hypothetical example of cost-benefit calculations from the three accounting perspectives, retaining employment training programs as the

example. The dollar figures in that exhibit are oversimplifications but nonetheless illustrate the different ways of calculating costs and benefits from the different accounting perspectives. Note, for instance, that the same components may enter into the calculation as benefits from one perspective and as costs from another, and that the difference between benefits and costs, or net benefit, will vary, depending on the accounting perspective used.

Exhibit 10-G Example of Cost-Benefit Calculations From Different Accounting Perspectives for a Hypothetical Employment Training Program

Benefits/Costs	Amount		
(1) Earnings improvement of trainees (before taxes)	\$100,000		
(2) Earnings improvement of trainees (after taxes)	80,000		
(3) Value of work done in training period	10,000		
(4) Project costs for facility and personnel	50,000		
(5) Project costs for equipment and supplies	5,000		
(6) Trainee stipends (direct transfer payments)	12,000		
(7) Earnings forgone by trainees (before taxes)	11,000		
(8) Earnings forgone by trainees (after taxes)	9,000		
(9) Taxes lost: (7) – (8)	2,000		
	Individual	Program Sponsor	Communal
Benefits	(2) 80,000	(1) – (2) 20,000	(1) 100,000
	(6) 12,000	(3) 10,000	(3) 10,000
	<u>92,000</u>	<u>30,000</u>	<u>110,000</u>
Costs	(8) 9,000	(4) 50,000	(4) 50,000
		(5) 5,000	(5) 5,000
		(6) 12,000	(7) 11,000
		(7) – (8) 2,000	<u>66,000</u>
		<u>69,000</u>	
Net benefit (benefits minus costs)	83,000	–39,000	44,000 ^a

a. Note that net social (communal) benefit can be split into net benefit for trainees plus net benefit for the government; in this case, the latter is negative: $83,000 + (-39,000) = 44,000$.

Note that net social (communal) benefit can be split into net benefit for trainees plus net benefit for the government; in this case, the latter is negative: $83,000 + (-39,000) = 44,000$.

The decision about which accounting perspective to use depends on the stakeholders who constitute the audience for the analysis, or who have sponsored it. In this sense, the selection of the accounting perspective is a political choice. An analyst employed by a private foundation interested primarily in containing the costs of hospital care, for example, will likely take the program sponsor's accounting perspective, emphasizing the perspective of hospitals. The analyst might ignore the issue of whether the cost-containment program that has the highest net benefits from a sponsor accounting perspective might actually show a negative cost-to-benefit value when viewed from the standpoint of the individual. This could be the case, for example, if the individual accounting perspective included the opportunity costs involved in having family members stay home from work because the early discharge of patients required them to provide the bedside care ordinarily received in the hospital.

Generally, the communal accounting perspective is the most politically neutral. If analyses using this perspective are done properly, the information gained from an individual or a program sponsor perspective will be included as data about the distribution of costs and benefits. Another approach is to undertake cost-benefit analyses from more than one accounting perspective. The important point, however, is that cost-benefit analyses, like other evaluation activities, have political features.

In some cases, it may be necessary to undertake a number of analyses. For example, if a government group and a private foundation jointly sponsor a program, separate analyses may be required for each to judge the return on its investment. Also, the analyst might want to calculate the costs and benefits to different subgroups, such as the direct and indirect targets of a program. For example, many communities try to provide employment opportunities for residents by offering tax advantages to industrial corporations if they build their plants there. Costs-to-benefits comparisons could be calculated for the employer, the employees, and also the average resident of the community, whose taxes may rise to take up the slack resulting from the tax break to the factory owners. Other refinements might be included as well. We exclude direct subsidies, for example, the transfer payments in the employment training example in [Exhibit 10-G](#), from a communal perspective, both as a cost and as a benefit, because they are

expected to balance each other out; however, under certain conditions it may be that the actual economic benefit of the subsidies is less than the cost.

Measuring Costs and Benefits

A particular challenge for cost-benefit analysis of social programs is identifying and measuring all the relevant components of program costs and benefits. When important benefits are disregarded because they cannot be measured or monetized, the project may appear less efficient than it actually is. If certain costs are omitted, the project will seem more efficient than it is. The results may be just as misleading if estimates of costs or benefits are either too conservative or too generous. These problems are most acute for ex ante analysis, in which there often are only speculative estimates of costs and impact. However, data may be limited in ex post cost-benefit analyses as well. The information from an evaluation of a social program may provide insufficient detail about the nature of the program and its effects to support a retrospective cost-benefit analysis. The analyst thus must frequently use additional sources of information or substitute informed judgments.

Monetizing Benefits

Social programs frequently do not produce results that can be easily valued in economic terms. For example, it may not be possible for the benefits of a suicide prevention project, a literacy campaign, or a program providing training in improved health practices to be monetized in ways acceptable to key stakeholders. What dollar value should be placed on the embarrassment of an adult who cannot read? In such cases, cost-effectiveness analysis may be a more appropriate alternative because it does not require that benefits be valued in terms of money, only that they be quantified by outcome measures.

However, because of the advantages of expressing benefits in monetary terms so that costs and benefits can be compared in the same familiar, meaningful dollar-value units, a number of approaches have been developed that may be applicable to the benefits produced by at least some social programs. The five that follow are frequently used.

1. *Money measurement.* The least controversial approach is to estimate direct monetary benefits when feasible. For example, if keeping a health center open for 2 hours in the evening reduces patients' absence from work (and thus loss of wages) by an average of 10 hours per year, then from an individual perspective the annual benefit of that particular influence can be calculated by multiplying the average wage by 10 hours by the number of employed patients.
2. *Market valuation.* Another relatively straightforward approach is to monetize gains or impacts by valuing them at market prices. If crime is reduced in a community by 50%, for example, one of the benefits might be an increase in the market value of the housing in that community. That benefit could be estimated as the difference between the housing prices before the decrease in crime and the housing prices in communities with crime rates comparable with those after the decrease and with similar social profiles.
3. *Econometric estimation.* Another approach is to monetize the value of a program effect by using a statistical model to estimate the independent influence of that impact on some domain of economic activity. For example, one of the benefits of reduced crime might be an increase in tax receipts from more business revenue. However, there are many other factors that influence business revenue. An econometric analysis might then be conducted with data on tax revenues and the factors expected to influence them from multiple communities with varying crime rates. That analysis would be structured to estimate the differential in tax revenues associated with different crime rates net of the influence of the other factors unrelated to crime rates that influence those revenues. The results could then be used to estimate the tax revenue benefit associated with the particular magnitude of the crime reduction effect of the program for which benefits are being monetized.
4. *Hypothetical questions.* A rather problematic approach sometimes used to estimate the value of intrinsically nonmonetary benefits is questioning the recipients of those benefits. For instance, a program to prevent dental disease may decrease participants' cavities by an average of one by age 40. That effect might be monetized by surveying participants about how much they think it is worth to have an additional intact tooth at that age or, perhaps, how much they would be

- willing to pay for that outcome. Such estimates are inherently subjective and somewhat speculative, and thus open to skepticism.
5. *Observing funding allocations.* Another approach is to monetize benefits on the basis of budgetary allocations by relevant social agents. For example, if state legislatures consistently appropriate funds for high-risk infant medical programs at a rate that works out to be \$40,000 per child saved, that figure could be used as an estimate of the monetary benefits of the effects of such a program on lives saved. Estimates may be similarly derived from the funding choices made by other program sponsors (e.g., foundations or businesses). Given that the process of making such budgetary allocations is generally complex, shifting, and inconsistent, this approach is necessarily rather tentative.

Estimating Costs

The most direct way to estimate program costs is to use the actual program expenditures for the various resources required to operate the program. The salaries of personnel, rents, payments for utilities, and other such direct expenses are typically represented in some form in a program's financial records. Extracting that information, however, may require digging into records on individual transactions in order to disaggregate the expenses summarized in broad categories in the program's financial reports. For instance, personnel costs may be a single line item in those financial reports, but the cost analyst may need to separate the costs for administrative personnel from those of line staff who work directly with program participants.

When direct expenditure data are not available, the analyst may turn to market price estimates for the cost of a particular program component. The market price is what a given program component would cost if purchased in the economic context within which the program operates. Suppose, for instance, that a program operates out of space donated by the organization that owns the facility in which that space is located. Though the program does not pay for that space, it is nonetheless a resource with value that is required to operate the program. The economic value of that space might

then be estimated on the basis of the average per square foot rental cost of comparable space in the community where the program is located.

Sometimes neither actual expenditures nor market prices represent the true value of a resource required to operate a program, or they are not available for that resource. The preferred procedure for estimating cost in those circumstances is to use **shadow prices**, also known as *accounting prices*. Shadow prices are derived prices for goods and services that are supposed to reflect their true economic value. Suppose, for example, that a program is located in a place where wages are artificially depressed, perhaps because of high unemployment or a depressed economy in an underdeveloped country. In such circumstances, the cost analyst may not believe that the actual wages paid to program personnel, or the wages for comparable personnel in the local market, represent the actual value of those personnel, that is, what their wages would be without the market distortions that suppress them. The analyst may then draw on whatever relevant information can be obtained to derive shadow prices for personnel costs that better estimate their economic value absent the local distortions.

Shadow pricing might also be used to value certain intangibles in the resources a program uses that are not easily captured otherwise. Suppose, for instance, that a number of university professors volunteer part-time to tutor children with reading difficulties. As a resource to the program, their economic value must be included in program costs. But, as volunteers, they are not paid, so there are no direct expenditures to account for. And these volunteers are not functioning as university professors in their program role, so the prevailing wages for professors do not provide relevant estimates. On the other hand, the program has the benefit of especially well-educated tutors, an intangible that may nonetheless contribute to the program's effectiveness. The analyst may then attempt to develop a shadow price for the time devoted by these volunteers using, perhaps, a wage rate estimate somewhere between the market price for professors and that for less highly educated but otherwise qualified individuals who could be hired to provide tutoring.

Another cost component often relevant for cost-benefit analysis of social programs is **opportunity costs**. The concept of opportunity costs follows

from recognition that individuals and organizations must choose how to allocate their resources from some set of reasonable and appropriate options. The opportunity cost of each choice is the value of the forgone options.

Although this concept is relatively simple, the actual estimation of opportunity costs is often complex. For example, a police department may decide to pay the tuition of police officers who want to go to graduate school in psychology or social work on the grounds that this additional schooling will improve their job performance. Given a fixed budget, however, the money used for those tuition payments is therefore not available for other uses. Although the tuition payments are a direct expenditure that can be accounted for, the cost of this program also includes the value of the loss to the department of whatever that money might otherwise have been spent on. The cost analyst must then try to make a reasonable determination of what those other uses would have been. Perhaps on the basis of a review of present and past budgets, the analyst decides that the primary adjustment has been to keep some of the police cars in service for an extra 2 months past when they otherwise would have been replaced. The opportunity costs of the tuition support payments might then be estimated as the additional repair costs that would be incurred during a 2-month extension. Because opportunity costs can only be estimated, as in this example, by making assumptions about the alternative investments, they are one of the controversial areas in cost-benefit analysis.

Other Considerations in Cost-Benefit Analysis

Secondary Effects (Externalities). Social programs may have secondary or external effects: side effects or unintended consequences that may be either beneficial or detrimental. Because such effects are not intended, they may be inappropriately overlooked in a cost-benefit analysis if an effort is not made to include them. Two types of such secondary effects are especially likely for social programs: displacement and vacuum effects. *Displacement* refers to program effects that push out something already in place in the program context. For example, a new publicly funded preschool program for 4-year-old children might displace programs run by community nonprofit organizations that cannot compete with a free program. If the

community programs serve a broader age range, say 3- and 4-year-olds, displacing them has the undesirable effect of reducing preschool options for the younger children.

Vacuum effects refer to gaps left in the social context of a program that result from the impact of the program on that context. For example, an employment training program may produce a group of newly trained persons who move from low-wage jobs to higher paying ones. Those individuals have thus vacated the jobs they held previously, leaving a vacuum that other workers might fill or, if the market does not supply those other workers, that might disadvantage the organizations that previously employed them. Such secondary effects may be difficult to identify and measure but, once found, should be incorporated into any cost-benefit analysis.

Distributional Effects. Distributional effects refer to the distribution of program benefits and harms across those affected by the program. Ideally, of course, a program would produce only benefits and no harm, but program effects can fall short of that standard and still be judged beneficial overall. One yardstick, which economists call the *Pareto criterion*, is that a beneficial program makes at least one person better off and nobody worse off. The distributional framework for cost-benefit analysis, however, is *potential Pareto improvement*, under which it is assumed that there will be gains and losses, but the gains must outweigh the losses. In the context of program evaluation and cost-benefit analysis, the distributional issue relates to the question of who gains and who loses. A program may have overall average positive effects on its intended outcomes, but that may mask a pattern in which those with the least need benefit the most while those in greatest need benefit least or not at all.

Because cost-benefit analysis involves program benefits in a very direct way, it may be important in some situations to incorporate distributional effects into the analysis. This is done by applying a system of weights whereby some program benefits are valued more than others, and/or benefits received by some groups or individuals are valued more than others. The assumption is that some benefits for some persons are worth more than others to the community, whether for equity reasons or for their

contribution to human well-being, and thus should be weighted more heavily in a cost-benefit analysis. Thus, if a home-visitation support program for teen mothers yields healthier infants (with reduced health care costs) and allows more part-time employment opportunities for the mothers (increasing income), it might be argued that even when these outcomes were monetized, benefits to infants with possible lifelong implications are more socially valuable than additional income for their mothers. Similarly, the effects of a remedial reading program on the most disadvantaged participants might be viewed as more valuable than those for the less disadvantaged participants. That value differential might then be carried into the weight given to the monetary values assigned to the benefits of greater literacy.

The weights to be assigned for these purposes may be determined by the appropriate decision makers, in which case value judgments will obviously have to be made. Weights may also be derived via economic principles and assumptions related to social well-being (e.g., what leads to greater economic efficiency). In any case, it is clear that weights should not be applied to a cost-benefit analysis of a social program without explanation and justification. An intermediate approach to considerations of equity is to first investigate whether there are differential program effects for different participant subgroups defined around characteristics such as need, relative disadvantage, minority status, and the like. If so, cost-benefit calculations could be done separately for each subgroup in order to make any differences transparent. This would allow smaller benefit-cost relationships to be identified and recognized as nonetheless worthwhile if they occur for the subgroups for which positive effects are viewed as especially desirable.

Discounting. Another major consideration in cost-benefit analysis relates to the treatment of time when valuing program costs and benefits. Social programs vary in duration and may produce benefits that endure or appear long after the intervention has taken place. Indeed, the effects of most programs are expected to persist for at least some time after participation ends. Costs and benefits occurring at different points in time must, therefore, be made commensurable by taking into account the time at which they are measured and valued. The applicable technique, known as **discounting**, consists of converting future costs and benefits to a common

monetary base by adjusting them to their present values. The present value of an expenditure that is made at some time past the start date for the analysis, for example, is less than the dollar value required at that time. This can be understood intuitively as the greater current burden of a payment that must be made today relative to one that need not be made until next year. Viewed in investment terms, it means that a dollar invested today in, say, a low-risk government bond will have grown in value by some later time. The present value of that later amount, then, is not the amount itself but the smaller amount one would have to invest now to be assured of having that later amount. Similar considerations apply to benefits. With the logic of “a bird in the hand is worth two in the bush,” a dollar’s worth of benefit in hand at the present time has greater value than the promise of that same dollar value at some future time.

Discounting in cost-benefit analysis, therefore, adjusts the dollar values of all future costs and benefits downward at some specified rate to transform them into present day values that are comparable irrespective of their temporal variation. [Exhibit 10-H](#) provides more detail about discounting and illustrates it with an example that shows how the applicable calculations are done.

The choice of time period on which to base the analysis depends on the nature of the program, whether the analysis is ex ante or ex post, and the period over which benefits are expected. There is no authoritative approach for fixing the discount rate. One choice is to set the rate on the basis of the opportunity costs of capital, that is, the rate of return that could be earned if the funds were invested elsewhere. But there are considerable differences in opportunity costs depending on whether the funds are invested in the private sector, as an individual might do, or in the public sector, as a quasi-government body may decide it must. The length of time involved and the degree of risk associated with the investment are additional considerations.

The results of a cost-benefit analysis are thus particularly sensitive to the choice of discount rate. In practice, evaluators usually resolve this controversial issue by carrying out discounting calculations with several different rates. Furthermore, instead of applying what may seem to be an arbitrary discount rate, the evaluator may calculate the program’s **internal**

rate of return, that is, the value the discount rate would have to be for program benefits to equal program costs. A related technique, *inflation adjustment*, is used when changes over time in asset prices should be taken into account. For example, the prices of houses and equipment may change considerably because of the increased or decreased value of the dollar at different times.

Earlier we referred to the net benefits of a program as the total benefits minus the total costs. The necessity of discounting means that **net benefits** are more precisely defined as the total discounted benefits minus the total discounted costs. This total is also referred to as the *net rate of return*.

Exhibit 10-H Discounting Costs and Benefits to Their Present Values

Discounting is based on the simple notion that it is preferable to have a given amount of money now than in the future. All else equal, current funds can be invested and earn compound interest that will make it worth more than its current face value in the future. Conceptually, discounting is the reverse of compound interest: It estimates how much we would have to put aside today to yield a fixed amount in the future. Algebraically, it is carried out by means of the simple formula:

$$\text{Present value of an amount} = \text{Amount}/(1 + r)^t,$$

where r is the discount rate (e.g., .05) and t is the number of years into the future at which the cost is incurred or the benefit is received. The total stream of costs and benefits of a program expressed in present values is obtained by adding up the discounted values for each successive year in the period chosen for study.

Suppose, for example, that a training program produces earnings increases of \$1,000 per year for each participant and the discount rate selected by the analyst is 10%. Over 5 years, the total discounted benefits using the formula above would be \$909.09 + \$826.45 + . . . + \$620.92, totaling to \$3,790.79, as shown in the table below. Thus, increases of \$1,000 per year for the next 5 years are not currently worth \$5,000 but only \$3,790.79. At a 5% discount rate, the total present value would be \$4,329.48. All else equal, benefits calculated using low discount rates will be greater than those calculated with high rates.

Year				
1	2	3	4	5
\$1,000	\$1,000	\$1,000	\$1,000	\$1,000
$(1 + .10)^1$	$(1 + .10)^2$	$(1 + .10)^3$	$(1 + .10)^4$	$(1 + .10)^5$
= \$909.09	= \$826.45	= \$751.32	= \$683.01	= \$620.92

Comparing Costs With Benefits

The final step in cost-benefit analysis consists of comparing total costs with total benefits. How this comparison is made depends to some extent on the purpose of the analysis and the conventions in the particular program sector. The most direct comparison can be made simply by subtracting costs from benefits after appropriate discounting. For example, a program may have costs of \$185,000 and calculated benefits of \$300,000. In this case, the net benefit (or profit, to use the business analogy) is \$115,000. Although generally more problematic and difficult to interpret, sometimes the ratio of benefits to costs is used rather than the net benefit.

In discussing the comparison of benefits and costs, we have noted the similarity to decision making in business. The analogy is real. In particular, in deciding which programs to support, some large private foundations actually phrase their decisions in investment terms. They may want to balance a high-risk venture (i.e., one that might show a high rate of return but has a low probability of success) with a low-risk program (one that probably has a much lower rate of return but a much higher probability of success). Thus, foundations, community organizations, or government bodies might wish to spread their investment risks by developing a portfolio of projects with different likelihoods and prospective amounts of benefit.

Sometimes, of course, the costs of a program or program practice are greater than its benefits. In [Exhibit 10-I](#), a cost-benefit analysis is presented that documents the negative cost-benefit relationship for the urine drug screening often required during hospital emergency care for patients then referred to psychiatric care. In this analysis, there were consequential costs associated with the drug screens, but no evident benefits.

Exhibit 10-I Cost but No Benefit From Emergency Room Urine Drug Screens

Substance abuse crises are common among those who visit hospital emergency rooms, often associated with preexisting psychiatric illness. These cases may then be referred to behavioral health services for further diagnosis and treatment. Many behavioral health centers require that a urine drug screen be completed and added to the medical records during the period of emergency care before these patients are transferred to behavioral

health. However, there is a cost associated with administering those drug screens, and they may extend the length of patients' stay in emergency care.

The authors of this study conducted a retrospective chart review for a sample of patients in a four-hospital community network who were transferred from emergency care to the psychiatric hospital in the network after evaluation and medical clearance. The sample consisted of 205 such patients who were discharged from the psychiatric hospital during a randomly chosen 1-month period. Clinical data were extracted and analyzed from the electronic medical record system for both the emergency care and the psychiatric services.

Of the 205 patients in the sample, 89 had a urine drug screen administered while they were in emergency care, and the remaining 116 did not. The records review revealed that the time to departure from emergency care was delayed for those receiving drug screens, but there were no other differences in the emergency care they received. Furthermore, the psychiatric care records showed no difference between patients with and without drug screens on the nature of the substance use disorders diagnosed, outpatient counseling or referrals for drug or alcohol counseling, or inpatient psychiatric hospitalization length of stay. Indeed, the drug screen results were not even mentioned in the psychiatric medical records for more than 75% of the patients who had received them.

The cost of the drug screen was estimated at \$235 per person, resulting in a total cost of \$20,915 for the 89 drug screens in the 1-month sample. Additional costs were associated with the extended time in emergency care for the screened patients, but those were not estimated. On the benefit side, the finding that the drug screens were not associated with significant differences in the emergency care provided, other than the drug screens, or with any differences in the psychiatric care provided, meant that no benefits were evident. The cost-benefit relationship, therefore, was negative: costs but no benefits. As the authors concluded, "Routine drug testing in stable psychiatric patients proved to be a waste of both time and money."

Source: Adapted from Riccoboni and Darracq (2018).

It bears mentioning that sometimes programs that show negative cost-benefit relationships are nevertheless socially important and should be continued. For example, there is a communal responsibility to provide support for severely disabled persons, but it is unlikely that any program that does so will have positive net value (costs subtracted from benefits) from either a program sponsor or communal perspective. In such cases, one may still want to use cost-benefit analysis to compare the efficiency of different programs even though none are expected to show positive net value.

When to Do Ex Post Cost-Benefit Analysis

A number of considerations are relevant to whether a cost-benefit analysis should be undertaken for a program that has already been implemented, that is, an ex post analysis. In some evaluation contexts, the technique is feasible, useful, and a logical component of a comprehensive evaluation; in others, its application may rest on dubious assumptions and be of limited utility.

Optimal prerequisites for an ex post cost-benefit analysis of a program include the following:

- The program has independent or separable funding: Its costs can be separated from those incurred by other activities.
- The program is beyond the developmental stage and there is reason to believe that its effects are significant.
- The program's impact and the magnitude of that impact are known or can be validly estimated.
- The benefits of the program can be represented in monetary terms.
- The results can be expected to interest decision makers who may consider alternative programs or whether to continue, expand, or revise the existing project.

Ex post efficiency estimation—both cost-benefit and cost-effectiveness analyses—builds naturally on the results of impact evaluations and adds a component of particular relevance to policymakers in circumstances in which consequential decisions about the program are at issue. [Exhibit 10-J](#) describes a cost-benefit analysis of that sort. An impact evaluation was conducted to estimate the effects of specialized treatment for violent juvenile offenders on their reoffense rates after treatment relative to practice as usual treatment. The differential cost of the specialized treatment relative to typical treatment was then compared with the differential cost of publicly funded criminal justice processing and prison costs for the subsequent arrests of each group. The results showed that the benefits of specialized treatment (cost savings) greatly outweighed the additional cost of that treatment.

Exhibit 10-J A Cost-Benefit Analysis of Specialized Treatment for Violent Juvenile Offenders

Serious juvenile offenders are typically sentenced by juvenile courts to some period of time in juvenile correctional facilities. Some youth do not do well in these facilities and are disruptive and aggressive in ways that do not support their own progress in the institutional treatment programs and can undermine the potential for successful outcomes by their less disruptive peers. In Wisconsin, the Mendota Juvenile Treatment Center (MJTC) is an alternative treatment facility designed to provide specialized mental health treatment to the most disturbed juvenile boys in the state's juvenile correctional facilities. In this study the impact of MJTC on postrelease delinquency relative to treatment as usual in the juvenile correctional facilities was evaluated. A cost-benefit analysis was then conducted to assess the cost of this specialized treatment relative to the monetary value of the reductions in subsequent offenses it produced relative to treatment as usual.

The intervention group in the impact evaluation consisted of 101 youth who were transferred to MJTC from two juvenile corrections institutions because of their disruptive and aggressive behavior. Using propensity scores based on a broad set of demographic, behavioral, and clinical variables, each of these youth was matched to a comparison youth who had been admitted to MJTC briefly for assessment or stabilization, then returned to the treatment-as-usual correctional facility for the majority of their treatment. Program effects were examined on three outcome variables assessed during a follow-up period of 53 months: all offenses, felony offenses, and violent offenses. That analysis found that the MJTC treatment significantly reduced the reoffense rates in all these categories. Youth in the matched comparison group averaged more than twice the number of charged offenses in the follow-up period on all these outcomes.

Cost calculations included only direct, tax-supported costs adjusted to 2001 dollars. For each participant, the cost of treatment in MJTC and the usual juvenile institution was calculated by multiplying the per diem cost by the number of days the youth resided in each setting. The cost for MJTC treatment per youth was \$161,932, which was \$7,014 more than the \$154,918 cost per youth for regular institutional treatment (an added cost of 4.5%).

Costs for the criminal justice processing of the postrelease offenses that constituted the program outcomes included the costs of arrest, prosecution, and defense as estimated from a national sample in other research plus the cost of incarceration for those who ended up in adult prison. The total of those costs over the follow-up period was \$11,080 per person for the MJTC treatment and \$61,470 for the comparison group, a \$50,390 difference favoring the treatment group. Thus, the additional cost of \$7,014 per person for MJTC treatment relative to treatment as usual for these difficult youth reduced their reoffense rates sufficiently to save \$50,390 per person in subsequent criminal justice costs, a savings of a bit more than \$7 for each additional dollar needed to cover the cost of the more specialized MJTC treatment.

Source: Adapted from Caldwell, Vitacco, and Van Rybroek (2006).

Conducting Cost-Effectiveness Analyses

Cost-benefit analysis allows evaluators to assess the economic efficiency of programs and compare the efficiency of alternative programs entirely in monetary terms. However, evaluators and stakeholders are often uneasy about cost-benefit analysis when applied to social programs. As we have noted, it can be difficult to obtain agreement on the monetary value of such outcomes as literacy, relief from depression, better marital relationships, or a teen suicide prevented. As a variation on cost-benefit analysis, cost-effectiveness analysis can be viewed as an informative alternative that does not require that such effects somehow be valued in dollar terms. Cost-effectiveness analysis is based on the same principles and uses the same methods as cost-benefit analysis, but it does not require that benefits and costs be reduced to a common monetary metric. Instead, the effectiveness of a program in reaching given substantive goals is related to the monetary value of the costs and efficiency is judged in terms of the costs for units of outcome.

Cost-effectiveness analysis thus yields estimates of the cost of obtaining the effects of the program represented in terms of the units on the respective outcome measure. Such estimates can then be compared across programs with effects on the same outcome to assess their relative efficiency. Cost-effectiveness analysis, then, is a particularly good method for evaluating the economic efficiency of programs with effects on similar outcomes without having to monetize the outcomes. Even without information about program effects, cost-effectiveness methods can be used to estimate costs per client served or similar unit costs, for instance, cost per treatment session (program outputs rather than outcomes), and compare those unit costs across programs providing similar services or with similar purposes.

Exhibit 10-K provides an example of a cost-effectiveness analysis of this sort for programs to increase high school completion rates among disadvantaged students at five different sites. Of particular interest to the evaluators was the variation across sites in the program cost per student and the even greater variation found in the cost per additional high school completer produced by the program at each site. The combination of differential program costs per student and differential impact on high school

completion resulted in wide variation in the cost-effectiveness of the different implementations of the program at the different sites.

Exhibit 10-K Wide Variation Across Sites in the Cost-Effectiveness of Support for High School Completion

Talent Search is a program to improve student progression through high school to college that has a long history in the United States. It is one of three educational outreach programs targeting students from disadvantaged backgrounds included in the 1965 Higher Education Act that was part of President Lyndon Johnson's War on Poverty. Talent Search is a large-scale program that, in 2011, provided services to 320,000 6th to 12th grade students from low-income homes designed to help them stay in school and on track for college. These services vary across sites, but may include counseling, informing students of career options, financial awareness training, cultural trips and college tours, help completing applications for student aid, preparation for college entrance exams, and assistance in selecting, applying to, and enrolling in college.

A critical prerequisite for entry into higher education is high school completion. A prior series of impact evaluations assessed the effect of Talent Search on high school completion, among other outcomes, in 15 Talent Search sites across Texas and Florida. Those evaluations used propensity score techniques to match Talent Search participants with students in the same high schools with similar rates of prior progression. These impact evaluations found that Talent Search participants outperformed the comparison group across all outcomes. For example, across the sites in Texas, 86% of the Talent Search participants completed high school compared with 77% in the comparison group, and in Florida, 84% of the participants completed high school compared with 70% in the comparison group.

Levin et al. (2012) were able to obtain cost data for five of the Talent Search sites included in the impact evaluations. At all of those sites the impact evaluation found that a higher percentage of Talent Search participants completed high school than comparison students, but there was considerable variation across sites, as shown in the table below.

High School Completion Rates in the Five Talent Search Sites Included in the Cost-Effectiveness Study

Site	Talent Search Participants	Comparison Students	Percentage Point Difference
A	90.4%	81.4%	9.1
B	88.3%	80.6%	7.7
C	63.4%	61.3%	2.1
D	96.7%	69.4%	27.3
E	85.0%	72.7%	12.4

To assemble cost data for each of these programs, all the cost components of each program were identified in the categories of personnel, facilities, materials,

transportation, and other. Items were included whether the program paid for them directly, they were paid from other sources, or they were provided in kind (e.g., facilities programs were allowed to use without payment). The price of each of these components was then estimated from a national price database the evaluators built for this project that included prices for more than 200 ingredients that might be used in an educational intervention.

These data showed considerable variation across the sites in the program cost per student. Combined with the estimates of the program's effects on high school completion from the impact evaluations, the cost associated with each student in the program who completed high school but would not have done so without program participation was calculated. Those results are shown in the table above.

Cost-Effectiveness of Talent Search on High School Completion at Five Sites

Site	Cost Per Participant	Total Number of Participants	Program Effect: Additional Completers	Additional Completers as Proportion of Total Number of Participants	Cost Per Additional Completer	Additional Completers Per \$100,000
A	\$4,900	615	56	.091	\$53,810	1.86
B	\$2,870	751	58	.077	\$37,250	2.68
C	\$2,770	1,100	23	.021	\$131,930	0.76
D	\$2,820	705	192	.272	\$10,330	9.68
E	\$3,650	759	94	.124	\$29,560	3.38
Overall	\$3,400	3,930	423	.108	\$30,660	3.26

This cost-effectiveness analysis revealed, first, that the per student cost of the Talent Search program varied widely across sites (from \$2,770 to \$4,900 per student), indicating that some were more efficient than others in providing their services. When the effectiveness of the programs are taken into account, there is even more variation in the cost per additional high school completer produced by the programs (from \$10,330 to \$131,930). Moreover, higher program costs were not closely related to either program effectiveness or cost-effectiveness. One of the sites with the lowest program cost (Site D at \$2,820 per participant) showed the largest program effect and, correspondingly, the lowest cost per additional high school completer produced.

Source: Adapted from Levin et al. (2012).

Assessment of the economic efficiency of social programs tops off the forms of evaluation that, altogether, constitute a comprehensive evaluation that addresses each of the critical domains of program performance. As depicted in this text, these include assessment of the needs a program aims to ameliorate, assessment of the program theory that articulates the program design and its rationale for addressing those needs, assessment of the implementation of the program defined by that theory, assessment of the impact of the program on the outcomes it intends to affect, and assessment of the relative costs and benefits of producing those effects. Although no single evaluation project is likely to encompass all these forms of

evaluation, all are relevant to some program circumstances and some stakeholder concerns and all have a place in the evaluation repertoire.

Summary

- Efficiency analyses provide a framework for relating program costs to program effects. Whereas cost-benefit analyses directly compare benefits with costs in commensurable monetary terms, cost-effectiveness analyses relate costs expressed in monetary terms to units of substantive effects achieved.
- Efficiency analyses can be useful at all stages of a program, from planning through implementation and modification. Evaluators distinguish between ex post analyses of the economic efficiency of programs already implemented and ex ante analyses of the expected economic efficiency of programs in the planning stage. Obtaining reasonably sound estimates of costs and benefits is more challenging before program implementation than afterward but nonetheless allows some systematic appraisal of cost and efficiency considerations to be included in program planning.
- Efficiency analyses make different assumptions and may produce correspondingly different results depending on which accounting perspective is taken: that of program participants, program sponsors, or the community. Which perspective should be taken depends on the intended consumers of the analysis and its purposes.
- Cost-benefit analysis requires that program costs and benefits be known, quantified, and transformed to common monetary units. Options for monetizing program effects (benefits) include money measurement, market valuation, econometric estimation, hypothetical questions, and observation of funding allocations. Shadow prices are used for costs and benefits when market prices are unavailable or, in some circumstances, as substitutes for market prices that may be unrealistic.
- There are various distinctive aspects of economic analyses that must often be taken into consideration. One of these is the concept of opportunity costs: the value of forgone alternatives to program involvement. Another is concern about secondary and distributional effects related to who benefits more or less from a program as a result of its intended and unintended effects.
- In cost-benefit analysis, when costs and benefits must be projected into the future, their monetary value must be discounted to reflect a common basis in present values.
- Cost-effectiveness analysis is a feasible alternative to cost-benefit analysis when benefits cannot be calibrated in monetary units. It permits comparison of programs with similar goals in terms of their relative efficiency and can be used to analyze the relative efficiency of variations of a program.

Key Concepts

Accounting perspectives 239
Benefits 238
Costs 238
Discounting 255
Distributional effects 246
Ex ante efficiency analysis 239
Ex post efficiency analysis 239
Internal rate of return 255
Net benefits 255
Opportunity costs 253
Secondary effects 246
Shadow prices 252

Critical Thinking/Discussion Questions

1. Compare and contrast ex ante and ex post efficiency analyses.
2. Explain how efficiency analyses can be useful at different stages in a program (planning through implementation and modification).
3. What are the basic procedures used in conducting a cost-benefit analysis?
4. Describe the five commonly used ways to express benefits of a program in monetary terms.

Application Exercises

1. Locate three cost-benefit analyses conducted on social programs and determine how the researchers monetized the benefits of the programs.
2. Design a short social intervention and then list the costs and benefits you would need to calculate to be able to determine the cost-effectiveness of the intervention.

Chapter 11 Planning an Evaluation

Evaluation Purpose and Scope

Research Questions

Research Design

Sample

Measures or Observations

Data Collection, Acquisition, and Management

Primary and Secondary Data Sources

Quantitative Data, Qualitative Data, or Mixed Data Sources

Administration of Data Collection: Primary Data

Data Acquisition and Database Construction

Data Analysis Plan

Communication Plan

Reports

Briefings and Interactions

Project Management Plan

Personnel

Resources

Study Timeline

Summary

Key Concepts

Preparing an evaluation plan is a necessity for all evaluations. An evaluation plan, which is often the culmination of extensive discussions about the goals, objectives, and methods for the evaluation, provides a document that will guide the evaluators conducting the evaluation. In addition, the plan sets the expectations for key stakeholders about their involvement in the process and the reports and briefings that will be produced. The plan defines the main purposes for the evaluation, the types of data and measures that will be obtained, the analyses that will be conducted, the resources to be allocated, how the project will be managed, and the means for communicating about the project and its findings. An evaluation plan guides the evaluators and also ensures that key stakeholders agree to similar expectations, including the communications about the findings.

The first 10 chapters of this book have been organized around the five domains of evaluations, which makes it clear that evaluations can serve many different purposes, from assessing needs and developing a program

theory to estimating impacts and calculating benefit-to-cost ratios. No matter which purpose or purposes have been chosen when an evaluation is being tailored to meet stakeholders' needs, an evaluation plan will be required to provide more details about how the evaluation will be carried out. Evaluation plans serve several useful functions. First, the plan describes the purpose for the evaluation, explicitly lays out the research questions that will be answered by the evaluation, and describes the research design, including data, measures, study sample, and analysis for the evaluation. Second, it describes the main evaluation activities, the timeline, and the resources that will be needed. Finally, it provides a common set of expectations for processes, procedures, and communicating findings for everyone involved in sponsoring and carrying out a particular evaluation. This final function is often overlooked when evaluations are being planned, but experience leads us to believe that it is extremely valuable to avoid misunderstandings when the evaluation is being carried out and finalized.

Evaluation plans can take many different forms. For evaluations conducted by independent evaluators who are external to the organization whose program is being evaluated, the evaluation plan will often take the form of a proposal or application. For large-scale evaluations of national, state, or other large-scale programs, the requirements for the proposals are often prescribed in great detail, usually to meet requirements for procurement of services or a grant application. In [Exhibit 11-A](#), the requirements of one agency that funds numerous evaluations, the Institute of Education Sciences in the U.S. Department of Education, are summarized. In other cases, the requirements for large-scale evaluations are specified by the sponsor to ensure that the evaluation will meet the needs defined by legislative bodies or other key decision makers. With smaller scale programs, such as those run by a local nongovernmental organization or a grantee of a philanthropic organization, or for less extensive evaluations the format may be less prescriptive and the plan less formal. However, even though the scope of both the evaluations and the programs to be evaluated may differ greatly, most evaluation plans have common components. Five separate but interrelated components for evaluation plans that will be the focus of the remainder of this chapter are (a) purpose and scope, (b) data collection,

acquisition and management, (c) data analysis, (d) communication, and (e) project management.

Evaluation Purpose and Scope

To begin to describe an evaluation's purpose and scope, every evaluation should be linked to one or more of the five domains of evaluation questions and methods, including needs assessment, assessment of program theory and design, assessment of program process, impact evaluation, and cost and efficiency assessment, which were described in the previous chapters. The order of the list of the evaluation domains does not imply that the evaluations should be undertaken in any particular order. It is not necessary to have conducted a needs assessment before measuring and monitoring program outcomes, for example. Also, evaluations can address questions that are raised in two or more of the domains. The purpose of any individual evaluation is selected primarily on the basis of the priorities of sponsors and key stakeholders and the evaluation's potential for influence at the point when the evaluation is being planned.

Influence can take many forms, including direct actions to change the program or changing attitudes about the program or its intended beneficiaries. Evaluations can be planned to influence *individuals*, usually the attitudes and actions of key stakeholders or decision makers; *interpersonal behaviors*, such as negotiations between program operators and administrators; or *collective actions*, including program adoption, improvement, expansion, or termination (Henry & Mark, 2003; Mark & Henry, 2004). It is generally recognized that the most common target for evaluations' influence is to improve programs, and this is particularly likely when the evaluation is sponsored by the agency administering the programs. Evaluations with program improvement as their primary purpose often include monitoring program processes and implementation. This may occur because the evaluation sponsors are also the program's administrators and they have a substantial interest in, as well as control over, increasing the quality or consistency of services, thereby improving the opportunities for the evaluation to influence decisions and actions.

Institute of Education Sciences Requirements for Impact (Efficacy and Replication)
Evaluations Grant Applications (Institute of Education Sciences, May 13, 2015)

Goal: Grant “supports the evaluation of fully-developed education interventions to determine whether they produce a beneficial impact on student education outcomes relative to a counterfactual when they are implemented under ideal or routine conditions by the end user in authentic education settings” (p. 45).

Required Component	Content to Be Included
Abstract (maximum length: one single-spaced page)	Title, topic and goal (from the lists provided in request for application), purpose, setting, population/sample, intervention, control condition, research design and methods, key measures, and data analytic strategy
Project narrative (maximum length: 25 pages)	Significance: description of the program/intervention to be evaluated, including problem it is intended to address, documentation that it is fully developed, and theory of change for the intervention
	Research plan: sample, setting, measures (outcomes, mediators and moderators, and fidelity of implementation and comparison group practices), research design, power analysis, data analysis procedures, cost analysis, and timeline
	Personnel: key personnel on the project team (4-page biographical sketches in appendix), management structure, plan to ensure objectivity, and contributions to prior evaluations
	Resources: capacity to manage the grant, access to resources relevant to conducting the evaluation, resources to be acquired to conduct evaluation, access to project sites (agreements with stakeholders in appendix), and access to secondary data (agreements with data sources in appendix)
Additional miscellaneous requirements	Data management plan Human subjects narrative Bibliography and references cited

Source: Institute of Education Sciences (2015).

Source: Institute of Education Sciences (2015).

But the program context and opportunities for influence could lead to very different choices about an evaluation’s purpose. As an example, Congress mandated an evaluation assessing the program impact of the national Head Start preschool program in 1998 to address two sets of questions:

“What difference does Head Start make to key outcomes of development and learning (and in particular, the multiple domains of school readiness) for low-income children? What difference does Head Start make to parental practices that contribute to children’s school readiness?”

“Under what circumstances does Head Start achieve the greatest impact? What works for which children? What Head Start services are

most related to impact?” (Puma, Bell, Cook, & Heid, 2010, p. i)

Although prior Head Start evaluations sponsored by the U.S. Department of Health and Human Resources had focused on program improvement and providing information on the children being served, Congress felt the need to have more definitive information on program impact. This need arose in part because of the dearth of existing information about the program’s impact on children’s readiness for school when it had been operating for more than 30 years at that time. To make decisions about the nature and level of continuing legislative support for Head Start, Congress demanded information about the program’s impacts. Although Congress expected a definitive evaluation of the program’s impact, there was also a realization that realistically it would take time for a high-quality evaluation to be conducted. Indeed, 12 years passed between issuing the mandate and producing the final report, although several interim reports on shorter term outcomes were available in the meantime.

Decisions about an evaluation’s purpose will be highly contextualized, as we noted in earlier chapters. The potential influence of an evaluation at any particular time will depend on several factors, including the developmental status of the program, the decisions that are most likely to be made about the program in the near term, and the purpose and findings of any prior evaluations that have been undertaken. For new programs, measuring the quality and implementation fidelity of a program’s services may produce the most immediately actionable evaluation findings. Often new programs have difficulties delivering services consistently and efficiently during their start-up phase, and findings about the quality of services or fidelity of implementation can be influential. In addition, an evaluation assessing program impact may be premature if services are not yet being delivered as intended or if the program’s targeted beneficiaries have not yet received a full “dose” of services. Needs assessments and evaluations focusing on program theory may be done even before programs are initiated. It may also be the case that the best time to begin to measure and monitor program outcomes and performance is before the program starts to actually deliver its intended services in order to have proper measures at the baseline before the program has had the opportunity to affect outcomes. But needs

assessments can also be conducted after programs have been operating for several years and systematic information is needed about any gaps between actual and intended outcomes or the extent of program coverage.

In addition, longer term and more comprehensive evaluations often may be called upon to serve more than one purpose. For example, in 2010, North Carolina won the competition for one of the federal Race to the Top awards and received \$400 million to establish and implement numerous statewide education initiatives as well as fund local initiatives intended to support the achievement of statewide goals in each of 115 local school districts and 28 charter schools (Fuller, Roy, Belskaya, & Leland, 2015). One of the features of the North Carolina proposal that was noted positively by the reviewers was that it included a comprehensive evaluation (to access the evaluation reports, see <http://cerenc.org/rttt-evaluation/>). In the first few months after the award, the team implementing the comprehensive evaluation developed and assessed the program theory for each statewide initiative as each was being formulated for implementation on the basis of the original program proposal. Throughout the first 4 years of implementation, the evaluation team assessed and monitored the implementation of each statewide initiative, its short-term performance, and the associated state and local expenditures. During the 5th and final year of the evaluation, the evaluation team estimated the impacts of the initiatives and projected the costs for sustaining the initiatives. As in this example, longer term, more comprehensive evaluations that serve multiple purposes may phase in different components over time on the basis of the programmatic context, including the maturity of the programs and their implementation status as well as to synchronize the availability of evaluation findings with the types of decisions that the evaluation may influence. For instance, in the first few years of the Race to the Top initiatives, program improvement decisions were expected to be most relevant, while later the state requested information of the impacts of the various initiatives to make decisions about which of them should be continued when federal funds were no longer available.

Once the evaluation purpose or purposes and its scope have been established, then the focus of the evaluation can be further described in the evaluation plan. Enumerating the research questions that will be addressed

and describing the research design and the measures or observations to be used help explain an evaluation's focus.

Research Questions

Developing the research questions allows the evaluator to precisely and succinctly specify the purpose of the evaluation and clarify how the evaluation will make judgments. In [Chapter 1](#), we listed common questions for each category of evaluation. Those examples of common questions are broad and general. In the evaluation plan, the questions are specific to the program to be evaluated and its objectives (see [Exhibit 11-B](#) for an example of the research questions for an evaluation of school turnaround in North Carolina). For each question, the measures or observations that will provide the evidence with which the program objectives will be assessed should be specified. Or, if the measures are too numerous, the types of measures that will be used in the evaluation should be listed. For example, the two questions addressed in the Head Start Impact Study (see above) described the outcome measures for the study as “multiple domains of school readiness,” which is understood by those working in this field to include social and emotional development, cognition and general knowledge, physical well-being and motor development, and approaches to learning. Thus, this research question indicates that the Head Start evaluation would use numerous and diverse measures to form a comprehensive picture of the program’s objectives for the well-being of the children served; the details for those measures were then provided later in the evaluation plan.

Exhibit 11-B Example Research Questions for a Comprehensive Evaluation

Research Questions for the Evaluation of the Transformation of North Carolina's Lowest Performing Schools	
1. Assessing overall impacts	What are the effects of North Carolina School Transformation on intermediate outcomes such as teacher mobility and instructional practices and student outcomes, such as achievement and disciplinary incidents?
2. Assessing subgroup impacts	What are the effects of North Carolina School Transformation on the outcomes of the lowest achieving students, defined as those not achieving proficiency on either mathematics or reading tests in the prior year?
3. Probing for underlying mechanisms or "active ingredients"	What underlying processes, such as student engagement or academic press, appear to mediate or suppress overall effects and effects on the lowest achieving students?
4. Assessing processes and implementation fidelity	What is the quality of the implementation of transformation services, including the comprehensive needs assessment, leadership coaching, instructional coaching, and professional development activities in turnaround schools? Are the transformation services implemented with fidelity? What is the amount and quality of comparable services delivered to control group schools during the study period?

In addition to specifying the outcome measures or the criteria that will be used to make judgments about the program's objectives, the research questions should focus the evaluation on the specific target population for the study. Assessments of program impacts may specify that the effects will be measured for the individual program participants or at the community level. For example, evaluations of cash transfer or contingent social benefits programs may look at the income, educational attainment, or risky behaviors of adolescents in the families that received cash payments (Heinrich & Brill, 2015) or at the average income of families in communities where the cash payments were provided, in an attempt to capture direct and spillover effects of the benefits (Diaz & Handa, 2006). Rather than defining the study population in terms of individuals or families that may benefit from program participants, evaluations assessing program processes may specify service delivery units (e.g., schools, health clinics), geographic areas, jurisdictions, or sites that will define the study units for the evaluation. Often research questions will begin with a phrase that clarifies the study population by reference to geography (youth in juvenile

detention facilities in Tennessee) or program participation (participants in the Scared Straight program in Los Angeles over the past 4 years). Also, in many cases, subpopulations of particular interest for the evaluation that will be the focus of special analyses, such as rural program sites or individuals from specific ethnic or racial minority groups, should be specified in the research questions.

Finally, the **standards** that the program is expected to meet are important components of the research questions where they are applicable. Standards can be empirically based within the evaluation, such as comparing the timing or quality of program service delivery with those of other programs providing similar services that have been judged to be of high quality or found to be effective. For example, an evaluation of a transitional housing program for victims of domestic violence may refer to studies of other providers of these services, such as *Transitional Housing Services for Victims of Domestic Violence: A Report From the Housing Committee of the National Task Force to End Sexual and Domestic Violence* (Correia & Melbin, 2005), to formulate standards for the program being evaluated. This report describes staffing caseloads for full-time caseworkers that could be used as a basis for objectives for the number of families that are active in the program and families with follow-up for each caseworker at any given time. Also, the norms of practice in other agencies with similar missions or the empirical standards they actually achieve can provide relevant benchmarks for assessing program performance. Alternatively, standards on specific criteria can come from authoritative sources such as professional organizations or legislation. For instance, the American Bar Association (2011) promulgates its *ABA Standards for Criminal Justice: Treatment of Prisoners*, which are standards for conditions of confinement and conduct and discipline that may be used for evaluating correctional programs. In some cases, standards can be generated from literature reviews conducted for the evaluation. In cases in which the evaluation literature on effective practices is too limited or thin for drawing conclusions about standards, expert opinions may suffice as the basis for standards.

During the planning process, adopting standards for an evaluation will usually require negotiation with key stakeholders and agreement before finalizing the plan. When an evaluation plan is prepared as a response to a

formal request, negotiations between the evaluators and sponsors may be precluded. In these cases, there is often a question-and-answer session with the sponsor and individuals interested in conducting the evaluation, but this may not allow an opportunity for negotiation. In other, less formal situations, the evaluation team and the sponsors should consider discussing the elements of the plan, especially the standards, the measures, and how the evaluators plan to reach judgments about program performance to ensure full understanding and agreement before the evaluation begins.

In summary, the research questions for an evaluation plan should be specific enough to provide a clear indication of the focus of the evaluation but not so detailed that they inhibit clear communication of the intent of the evaluation. Finding the right balance will require both artfulness and craftsmanship from the evaluator. A main goal for the research questions is to be specific enough for stakeholders to be able to unambiguously determine whether the questions have been adequately addressed when the evaluation is completed. And in the interim, the questions will provide important guideposts for ensuring that the evaluation does not drift from its original intent as operational decisions are made. Usually, drafting these questions will fall to the evaluator, although sometimes they are specified in requests for proposals or evaluation tenders. However, during the process of finalizing the questions, evaluators should want, expect, and seek input from key stakeholders, and in some cases members of the communities, who are likely to be affected by the evaluation and its findings, to achieve a high degree of clarity about the meaning of the questions and credibility of the findings among all stakeholders.

Research Design

The purpose of the research design in the evaluation plan is to clarify the intent of the evaluation and provide an introductory overview of the sample, measures, data, and analysis for key stakeholders. Evaluations generally involve one or more of three types of research designs: descriptive, causal, or case study. Often case studies are combined with one of the other designs to provide complementary information, but are sometimes used effectively on their own, especially for development of detailed program theories or for smaller scale evaluations with limited resources.

Descriptive designs have the primary goal of providing sufficiently precise and accurate information, such as averages, percentages, ranges, or distributions of key study measures for a subset of the population (sample) that adequately represents the target population for the evaluation.

Descriptive designs are commonly used for all types of evaluations that involve empirical investigation, including needs assessments, process monitoring, and outcome monitoring, and may be combined with a causal research design for impact evaluations. In the research design section of the plan, it is important to orient readers to the domains or constructs that will be measured within categories, such as outcomes, characteristics of the population receiving services, or key process variables including amount of services offered and amount actually received. Also, the main sources of these data, such as administrative records, surveys, interviews, focus groups, or direct observations, should be described. In addition, a brief overview of the nature of the study population or sample should be included. The main analytical techniques are also frequently listed.

Causal designs are used exclusively for evaluations assessing program impact, although impact estimates may also be needed for assessments of cost and efficiency. Many of the elements of descriptive designs—measures, study sample or population, data sources, and analytical techniques—are similar to those required in evaluation plans for causal designs. The key difference is the addition of a plan for how the causal impact estimates are to be produced. Recall that the goal of causal designs, including randomized designs, regression discontinuity designs, and all the varieties of comparison group designs, is to isolate the effects of the

program on outcomes of interest to key stakeholders and provide unbiased estimates of the magnitude of the effects. In the description of the causal design in the evaluation plan, it is important to explain the specific design that will be implemented and identify its strengths, the threats to its validity and any other limitations, and the means by which potential biases will be mitigated. These designs and the interpretation of their findings are extensively discussed in [Chapters 6 to 9](#) and therefore will receive little additional attention here.

Case studies can be useful designs in many types of evaluations, and can be especially useful for developing program theories. In a recent study of school reform processes, for example, Thompson, Henry, and Preston (2016) selected 12 high schools as case study sites on the basis of the change in their rates of student proficiency on statewide tests during the implementation of the reform: 4 that had improved proficiency by 25 percentage points or more, 4 that averaged gains of about 15 percentage points, and 4 that had either worsened or improved by less than 5 percentage points. Then observations, interviews, and focus groups were conducted to contrast what had gone on during the implementation of the reform in these different sites and generate working hypotheses about the conditions and activities that led to successful school reform.

In this example, the contrasts were based on changes in the performance of the selected schools. In other case studies different contrasts may prove useful, such as on key process or outcome variables across localities (e.g., urban, small town, suburban, rural), service units of different sizes (e.g., number of personnel, size of budget), nature of the populations served (e.g., ages, racial/ethnic composition), or nature of the service delivery unit (e.g., jails, juvenile corrections centers, state correctional facilities, federal correctional facilities). In the plan for an evaluation using case studies, the rationale as well as the specific method for selecting the cases and the number of cases should be explained.

Sample

The purpose of the section of an evaluation plan about the sample is to describe the units to be selected for the study and how they will be selected. The primary goal of sample selection is to provide unbiased and sufficiently precise information on key study measures that adequately represents the target population for the evaluation. A sample is simply a subset of the units that make up the target population for an evaluation. To describe program processes and implementation, sites or service delivery units may be the study population to be sampled. For outcome monitoring, which by definition entails measures of the intended program beneficiaries, evaluators may plan to sample individuals, families, households, or communities to be served.

The highest level for representation of the target population comes from inclusion of the complete population in the study sample, for instance, when administrative data are available for measures of interest. Carefully drawn probability samples can also provide representative data, such as a random selection from a list of the target population for the study. Probability samples are prized in evaluations because the sample summary statistics, such as the mean, standard deviation, or interquartile range, can be generalized to the population from which the sample was drawn. In some cases, the sample will represent the full population, but only for a specific period, such as all children in foster care in Washington, D.C., during a specific year that may be presumed typical or the most current period available.

All other things equal, larger probability samples produce more precise estimates of the population, but even small probability samples have the benefit of being selected objectively. Eliminating human discretion can be especially important for evaluations because some stakeholders who are motivated to put the program in the best possible light may suggest collecting data from sites that may be operating most smoothly or individuals who are known to have benefited from the program. Human discretion can also work in the opposite direction if some stakeholders are critics of the program and are inclined to steer attention to poorer

performing sites or less successful participants. A major benefit of probability sampling is elimination of any such bias in the selection of the sample.

Probability sampling can be a complex undertaking, and a complete guide is beyond the scope of this evaluation text, but the topic was addressed more extensively in [Chapter 2](#). More detailed accounts can be found in such full-length volumes as Henry (1990) and Fowler (2014). In many cases, however, evaluators have such limited resources for data collection that probability sampling can be infeasible, or the data for the evaluation are intended to be exploratory and do not need to be generalizable to the target population. In such situations, evaluators often use convenience samples that are drawn to include typical cases or perhaps to maximize variability or heterogeneity. In contrast to probability samples, these nonprobability samples involve some amount of human discretion in the selection of the individual units. The selection process for nonprobability samples limits the generalizability of the data and statistics generated from those data. In planning for nonprobability samples, careful consideration is needed about the likely variations in the target population on key process or outcome variables across sites, operational units, population subgroups, and the like. Including these sources of variation in the evaluation sample may lead to more credible and justifiable representation of the individuals or service units that contribute to the evaluation. A primary consideration can be to obtain a sample that reflects the variation in the key characteristics that exist in the target population to ensure that the full range of program and participant differences are represented to at least some degree in the data on which the evaluation conclusions are based. In addition, evaluators may enhance the utility of the findings if data can be provided in categories useful to stakeholders, such as by administrative units, which implies that a sufficient number of units must be selected from each category to provide a meaningful description of that category.

For any evaluation that relies on sampling, important aspects of the sample selection to be described in the plan are the target population definition, operational definition of the target population (actual study population for the evaluation), size of the sample to be selected, and the method of selecting the sample. Additional information in the plan that is likely to be

useful for interpreting the findings has to do with missing data. Data can be missing for many reasons, including lack of cooperation of individuals selected in the sample as survey respondents or incomplete administrative data files. When describing the sampling procedures, the nature, number, and types of cases that may be intentionally or unavoidably excluded from the study samples should be noted. Also, the evaluator will need to consider if the amount of missing data, most usually from nonresponse, will require that a larger initial sample be selected (oversampling) to compensate for the missing data and allow a sufficient number to remain to support the planned analyses.

Measures or Observations

Listing the primary study measures is essential for explicating the focus of the evaluation. Here it is important to keep in mind that most measures that are useful for evaluative purposes will have an explicit and commonly agreed upon valence. More educational attainment is good—it has a positive valence. More recidivism of former inmates is bad—it has a negative valence. These are both outcome measures, but the principle applies not only to measures of outcomes but to measures of process or needs. For example, more time to process a case or provide services to clients who have been determined to need them has a negative valence. Recalling that an essential characteristic of an evaluation is making judgments about programs, purely descriptive measures without an explicit valence should be identified as such. In many cases, measures without explicit valence may be needed for context and description. For example, in impact evaluations that rely on covariates to adjust for differences between the program participants and those in the comparison group (discussed in [Chapter 7](#)), descriptions of those covariates and their intended role would be included in the measurement section of the evaluation plan, but they would not necessarily have any valence for evaluative purposes.

In many evaluations, the measures of program exposure will need careful consideration. The most obvious cases are those where evaluators intend to assess the associations between variations in program exposure and variations in outcomes. In these evaluations the duration of the treatment period for an individual, the amount of treatment, or the actual time participating in treatment-related activities can be used as measures. In a recent study of the effectiveness of cash transfer payments in South Africa, Heinrich and Brill (2015) examined the effects of dose and timing of cash transfers on adolescent participation in risky behaviors and educational attainment. A dose measure in this evaluation was the number of months that the adolescent received cash benefits. They found that many of the adolescents experienced periods when the receipt of benefits were interrupted, in most cases because of bureaucratic red tape. The actual dose was computed for adolescents by calculating the months between the start and stop dates for each adolescent and subtracting the periods when benefits

were interrupted. They found, for instance, that female adolescents who experienced interruptions in receipt of benefits were more likely to participate in criminal activity, were more likely to have had sex, and completed fewer grades of schooling. Measures of program exposure also may be taken at the service provider level rather than the individual participant level and may focus on completion of the intended service regimen or the access to service made available by the provider.

Another consideration for the selection of measures has to do with the level of subjectivity of the measures, that is, the extent to which the resulting data can be influenced by those collecting the data or the stakeholders involved, such as program personnel. Three distinct categories of measures based on the levels of subjectivity can be identified. First are the most objective (or least subjective) measures, which are made directly on the basis of behaviors, documents, or other artifacts, such as patient records or reports about interactions between program personnel and program participants, or alternatively, direct assessments or observations by evaluators using structured, systematic procedures. For example, in studies of interventions with young children, objective measures of children's developmental outcomes may be taken by trained, independent assessors using validated assessment instruments, such as the Woodcock-Johnson Tests of Cognitive Abilities. Alternatively, objective measures may come from administrative data sources for outcomes, such as periodic mental health status measures, or process variables, such as length of time between therapeutic sessions. When measures are actually used for administrative purposes, these data can be both complete and accurate for use in an evaluation.

A more subjective measure would involve measures of the same skills or behaviors, but based on survey responses or less formal assessments by knowledgeable informants. Continuing with the early childhood example, caregivers may be asked in a survey to estimate the number of letters a child can recognize or to recall the frequency with which someone reads to the child. Often these measures ask for specific counts or ranges of specific counts (e.g., how many letters of the alphabet a child can regularly recognize), or they use nonspecific quantifiers (e.g., *almost always*, *usually*, *occasionally*, *almost never*). These sorts of measures may be both less reliable and subject to systematic bias. Although they are more subjective

than direct assessments or observations, the evaluators make evaluative judgments on the basis of the responses, which is not the case with the most subjective measures described below.

The most subjective measures ask participants or knowledgeable informants to make judgments about the skills or behaviors of the intended beneficiaries of the program or rate the quality of program services. For example, caseworkers in a residential program for formerly homeless individuals may be asked to rate their readiness for independent living, a question that is inherently somewhat speculative. Other, often used examples of this type of measure are asking program participants about their satisfaction with services or whether they would recommend the program to someone in similar circumstances. In these cases, the respondents are making evaluative judgments directly. When describing the key measures in an evaluation plan, it is important to explain how the measurements will be taken and convey their level of objectivity. Stakeholders, especially evaluation sponsors, need to understand the objectivity of a proposed measure and may support the need for more objective measures for the evaluation to be credible, even if these measures cost more to collect than the more subjective measures.

Also, evaluation plans usually divide the key measures into categories on the basis of the purpose for which they will be used in the evaluation. For example, the most important category of measures for evaluations assessing effects or impacts is outcome measures. Outcome measures may be further subdivided into the more proximal and the more distal outcomes. Another type of evaluation, assessing and monitoring processes will likely focus on measures of process quality, implementation fidelity, program participation frequency (amount of time spent receiving services), and program dose. Between process and outcome measures lies a group of measures often called program outputs or activity measures. Unlike outcome measures and some of the process measures, outputs are usually measured at the program, site, or service delivery unit level, the analogue to McDonald's number of hamburgers served. These may be important for some evaluations to determine if the service agencies or organizations attained the reach in delivering services that they were expected to have. These measures can be useful for holding the service units accountable for the services they were

expected to deliver as well as for understanding and interpreting outcomes. Usually other measures such as program or site characteristics and participant characteristics are also listed in the evaluation plan in order to describe the units and the extent to which they vary on these measures.

Finally, evaluators planning an evaluation often generate a more extensive list of measures that mix in measures of outcomes or processes that the stakeholders need-to-know with nice-to-know measures. Nice-to-know measures are interesting to certain stakeholders, or the evaluators, but do not tie back directly to answering the research questions. They may allow comparisons of subpopulations of the targeted beneficiaries or add a different measure of the services but not add information that will be useful in addressing the key evaluation research questions. Often obtaining the measures that are most needed, especially from interviews or surveys, will necessitate carefully combing through the list of measures and determining a manageable number given the time that respondents can reasonably be expected to spend providing data.

Data Collection, Acquisition, and Management

Data collection and management activities often require the lion's share of the time devoted to conducting an evaluation. Even evaluations that rely primarily on secondary data sources often require that considerable time be spent on developing data sharing agreements; putting human subject protections in place; cleaning, merging, and managing databases; maintaining data security; and dealing with missing data. Those evaluations that involve primary data collection at sites that require travel and interactions with multiple participants and/or program personnel can require months of time and large numbers of staff hours to obtain permission to interact with human subjects, identify sites, obtain permission to collect data at each one, schedule visits when the key respondents are available, hire and employ the right number of evaluators with the qualifications and skills needed for the data collection, train evaluators who will collect the data, collect the data, transcribe or otherwise get it into proper form for analysis, and complete site visit memos. A detailed data collection and management plan helps ensure that the timeline for the evaluation and the resources required will match the level of effort needed to enact the plan.

Primary and Secondary Data Sources

For the purposes of the data collection part of the evaluation plan, it should begin by identifying a source for each key measure. The main distinction between data sources is whether they are primary or secondary sources. Evaluators themselves conduct **primary data** collection activities for the purposes of that specific evaluation. These activities usually consist of observations, interviews, focus groups, surveys, or direct assessments. The source for each usually includes the type of respondent, for example, program participants or caseworkers, or the locus of the data collection, for example, direct observations during delivery of services, and the method used for data collection from the list above. A single source and data collection method can, of course, provide numerous key measures. Often several of these data collection activities are bundled into “site visit” activities, sometimes referred to as protocols, to achieve the greatest efficiency possible, minimize travel, and reduce the burden on respondents. For evaluations in which on-site data collection is required, key measures should be listed under each specific data source, which includes the type of respondent and the data collection activity. For example, a data source could be corrections officer focus groups or superintendent interviews. The practice of listing each key measure with their data source helps ensure that data for all key measures are being collected.

Secondary data activities are those that begin by acquiring data from another source that was originally collected for purposes other than the evaluation at hand. Administrative databases are becoming very valuable secondary data sources in many evaluations. When data are actually being used for programmatic purposes, such as making payments for delivery of services or as a basis for deciding whether potential clients are assigned to treatment, the data are highly likely to be accurate and unlikely to be missing. For example, the administrative data source in which teachers’ years of experience are recorded to determine the salary payments for individual teachers is likely to be a very accurate measure of experience, which may be a key process measure for an evaluation of an education program. In other cases, data from secondary sources that are not being used for management or other purposes, such as items on intake forms that

are not used for deciding program eligibility or assignment, may be too frequently missing to be useful or in other ways of limited usefulness for the evaluation. During the planning stage, it is always prudent for evaluators to obtain a sample from the secondary data source they expect to use in the evaluation, deidentified if appropriate and necessary, to examine its completeness and whether the variable values are within the range of possible values. This check for missing data and out-of-range data reduces the possibility that key measures that are listed in codebooks or data dictionaries for the secondary data will not actually be available or useful from these sources. If the data are very important for the evaluation and not actually available or accurate from the secondary source, primary data collection may be needed to collect them, and to the extent possible, this should be known during the planning process.

Quantitative Data, Qualitative Data, or Mixed Data Sources

Historically, one of the most extensively debated issues in evaluation focused on the nature of the data to be collected: whether quantitative data, qualitative data, or mixed types of data should be collected. In part, the debate stemmed from philosophical differences in the evaluation community (see Mark, Henry, & Julnes, 2000, for a brief account of these differences) and, in part, it stemmed from pragmatic differences. We will focus on the pragmatic differences. Collecting quantitative data requires extensive planning and a significant investment of time to identify the units from which data will be collected; obtain permissions; review the literature to find measures used in prior studies; develop new measures when needed; combine the measures into instruments; pilot-test the instruments for cognitive load, respondent burden, reliability, and validity; administer the instruments; and compile the data for analysis. The goal of quantitative data collection is that the data be both valid and reliable. *Validity* refers to the accuracy or truth value of the data. In other words, the evaluators will want to know, do the data collected actually measure the constructs that were intended to be measured? Issues associated with validity are quite complex but essentially concern the truth value of the measures. *Reliability* refers to the consistency of the data that are collected. Would two individuals with the same attitude or behavior choose the same response on the measure? The quest for valid and reliable data places a premium on (a) finding measures that have been shown to be valid and reliable in prior research and (b) implementing procedures for collecting data that are independent of the individual actually collecting the data.

Qualitative data collection, for the most part, allows differences in the data collected on the basis of the experiences of the individuals from whom it is collected and the human agency of the data collector. The goal in much qualitative data collection is to adequately represent the experiences of those involved with the program from their own perspectives. In practice, evaluators can begin the collection of qualitative data early in the evaluation and use their interactions with program personnel and the program

participants to refine their data collection as they learn more about the program and the experiences of relevant stakeholders. This data collection is more flexible and adaptive than quantitative data collection, which as described above is more rigid. Working hypotheses based on the qualitative data collected at one site can be tested directly when collecting data at other sites. This type of research design is sometimes referred to as emergent and takes advantage of the capacity of the individuals collecting data to learn and modify their data collection activities on the basis of what they learn.

In many evaluations, both qualitative and quantitative data are collected, which are referred to as mixed-methods designs. Generally, mixed-methods designs attempt to leverage the strengths of both types of data collection for the evaluation in order to fully integrate the different types of information each provides and expand the utility of the evaluation findings (for more details, see Burch & Heinrich, 2016). Sometimes both types of data collection approaches go on concurrently, and in other cases, they are staged sequentially so that findings from one can be used to guide and inform the collection of a different type of data later. One popular mixed-methods approach begins with qualitative data collection and uses the findings to identify and prioritize the quantitative measures to be collected later. In other cases, the quantitative findings from the first phase of a study are used either to identify sites, for example, high-performing and low-performing sites, for qualitative data collection to understand key programmatic differences between the sites or to have participants interpret the findings from their own perspectives and experiences.

As a practical matter, the decision about quantitative or qualitative data or both should be based on the research questions, the type of data needed to guide or inform the decisions to be made about the program, the amount of time available for the evaluation, and the other resources available, including staffing and funds. In general, when more controversial programmatic issues and when individuals who are less directly involved with program operations are involved in the decisions that the evaluation is intended to influence, quantitative data will be highly valued. Often these are policy evaluations or large-scale program evaluations. In contrast, when those with direct involvement with the program are involved in using the evaluation findings, qualitative data may be more persuasive. These are

often smaller scale or local evaluations, and this distinction led Rossi (1994) to suggest that there was a natural division of labor for evaluators based on the scale of the policy or program and the scope of the evaluation.

Administration of Data Collection: Primary Data

Managing the data collection process for an evaluation can be one of the most logistically challenging activities for evaluators. This process begins once the decisions about the study sample and measures have been made and the instruments have been developed. Usually, the first step is to develop the protocols for data collection and along with the instruments submit them for review by an institutional review board that oversees research involving data collection that includes human interactions or identifiable data. Large research organization and universities have institutional review boards that evaluators employed by those organizations use for the reviews. Since the 1980s, independent institutional review boards have sprung up that provide reviews for researchers and evaluators who are not affiliated with a university or large research organization. More information about independent institutional review boards can be found on the Web site of the Consortium of Institutional Review Boards (<http://www.consortiumofirb.org>).

Another important step in the data collection process is obtaining permission to collect data at each site that has been chosen. When the data collection sites for an evaluation are part of numerous sites operated by a larger organization, such as regional employment offices operated by a statewide employment services agency or police stations that are units within a metropolitan police department, permission can be a two-step process. First, the evaluators will need to contact all of the large organizations within which they want to collect data to find out if they have a formal procedure for reviewing research in the organization and, if not, to whom the requests to collect data at each of the sites should be directed. If a formal process is in place, the evaluator will need to make a request to the individual or unit with the authority to review research proposals, such as the research director for the regional employment agency or the police department's research committee. The review process is often quite involved and requires much more time than novice evaluators may expect. For example, some organizations funnel all evaluation and research requests through an internal research review committee, and in many of these organizations the review committees meet only quarterly. If the request

arrives just after a review committee meeting, up to 90 days may pass before the next review opportunity.

An important consideration during the review process is having a plan when organizations or sites refuse to participate. If the evaluator plans to select the study sample sites using a probability sampling process, one of the best ways to handle refusals is to have oversampled the study population initially. Alternatively, after the main study sample has been selected, the evaluator could select another small random sample as a holdback sample. The size of the holdback sample is based on the number of refusals that are expected (perhaps prior experience suggests that 10% of the main study sample might refuse), and that number determines the size of the holdback sample. If the decision is made to replace any of the refusals with the holdback sample, then all of the holdback sample must be used to maintain a probability sample. Additional holdback samples can be selected, without replacement of any of the units previously selected, if needed. For nonprobability samples, replacements can be selected from the remaining sites that were not selected in the first round. If the main study sample was selected on the basis of inclusion criteria for producing an intentionally heterogeneous study sample, the sites that refused to participate could be replaced using the same criteria as the site that refused to participate, if any sites with those characteristics remain after the original sample was selected.

Also, before actual data collection begins, the site visits must be scheduled, data collectors with the skills needed for the data collection must be hired, and they must be trained on the protocol for the visits. Usually data collection follows a standard protocol that sets the length of the visit and each data collection activity, the number of participants, and either specific individuals or the types of individuals for each data collection activity (i.e., each focus group, interview, observation, or direct assessment). It is usually best to coordinate the visit with one individual at each site, start the process early enough for them to arrange appropriate participants and location for each data collection activity, and offer the site as much flexibility concerning the timing of the data collection activities and participants as possible. If the data collection needs to take place during particular events, say direct observations of training activities, obviously that will constrain

flexibility. Also, it may limit the times when the data collection can occur as well.

Finally, when the data have been collected, it is important to ensure that all instruments, notes, recordings, documents or other artifacts, and summary memos that are a part of the data to be collected have indeed been submitted by individuals responsible for data collection at each site and obtained by the individual responsible for overseeing the field work and processing the data. For larger evaluations, the latter is often the responsibility of a specific individual who is assigned to this task. No matter who is responsible for overseeing this, it is important to make sure the data collection team members for each site understand that it is their responsibility to produce all of the data required for the evaluation for that site. In addition, the documentation and original copies of responses, notes, and recordings must be stored in a manner that facilitates efficient access should questions or concerns arise during the analysis.

Collecting original data for any evaluation takes considerable time. This should be evident from this brief description of the processes. However, the time period allocated for actual data collection—site visits, surveys, or interviews—can stretch out for months. For example, the increasingly popular mixed-mode surveys, which are useful in many evaluations, often use responsive or adaptive designs to reduce nonresponse errors (Dillman, Smyth, & Christian, 2014). These designs require additional steps to identify and communicate with the types of individuals who have lower response rates in the earlier rounds of administering the instrument. In these cases, plans must include sufficient time to match the responses with existing data on respondent characteristics, estimate response rates for various groups, and develop a particular strategy to communicate with them, for example by phone versus e-mail, which prior research indicates may increase response rates (Dillman et al., 2014). When the time and effort for these follow-ups are not well planned in advance, the response rates for the surveys may not be adequate, and the resulting data may be biased because of nonresponse. Often the data collected when response rates are low are from those for whom the survey is most salient. Response rates that are very low may not be sufficient for stakeholders to consider the data collected to be credible and threaten the validity of the data for use in

the evaluation, especially when intended to represent the entire study population. Expertise in the type of data to be collected, the amount of time needed for appropriate follow-up to be implemented, and the timing of these activities to avoid burdening respondents during especially busy or stressful times of the year will be needed for the planning to increase the likelihood that sound data will be available for the evaluation.

Data Acquisition and Database Construction

The parallel process to administration of the data collection for primary data is data acquisition and database construction for secondary data. Database construction, although very different from original data collection, requires significant planning for a successful evaluation. Also, it is required for evaluations that intend to use both primary and secondary data and even for some evaluations that rely entirely on original data collection. The process usually begins with negotiating a data sharing agreement or memorandum of understanding with the individuals in the organization that has the data who have the authority to make the data available. Often when these are administrative data, the organization will have a set protocol for the data sharing agreement. Usually the agreements define the parties to the agreement; state the mutual benefits from the data sharing; define the purpose for the data sharing; list pertinent laws or statutes that govern the conditions for data sharing, such as the Family Educational Rights and Privacy Act of 1974 (FERPA) or the Health Insurance Portability and Accountability Act of 1996 (HIPPA); assign responsibilities to each of the parties for abiding by the legal provisions, transferring, maintaining, and protecting the security of the data to meet the legal and other requirements; describe the process for handling data requests; clarify who owns the data; explain the nature and extent of intellectual property rights of the parties receiving the data; provide permission to use the data and restrictions for its use; and list the provisions for handling disagreements between the parties. Both the administrative agency that has collected the data and the evaluator may benefit from being explicit about restricting the evaluator from transferring or otherwise sharing the data with any other party. In some cases, key evaluation stakeholders may wish to gain access to the data for other purposes, and a restriction in the data sharing agreement can redirect any discussion of data access to the original owner of the data, the administrative agency, and the stakeholder.

Once the data sharing agreements are in place, the processes of transferring the data, maintaining the data in a secure environment for both storage and analysis, cleaning the data, merging original data sets for analytical purposes, managing the analytical data sets, and dealing with missing data

require refined skills that will be needed on the evaluation team. The plan for the evaluation should ensure the availability of personnel with data management and security skills, time for carrying out these processes, and the facilities and software needed for the processes.

Data Analysis Plan

The plan for analyzing the data represents the penultimate step in an evaluation plan. It is a useful practice to organize the data analysis plan by each research question. Checking the alignment of research questions with the analysis plan by using the questions to organize the analysis can aid in the identification of research questions that have been omitted but seem important for the evaluation. For example, evaluators may omit some important descriptive information from the research questions that are important for context as well as aiding the interpretation of the findings. Because it takes time and expertise for the evaluators to answer these descriptive questions along with the highest priority research questions, it is important to include them in the analysis plan.

The data analysis plan will be quite different for qualitative and quantitative data. This is especially true when the qualitative data have been collected following an emergent design, which allows modifying the data collection procedures during the data collection process. Emergent designs will embed the initial part of the analysis during the data collection and require a process that iterates between data collection and data analysis as well as a strategy to communicate with all of those involved in data collection about potential modifications of the data collection protocols. But even with qualitative data collection plans that follow a set protocol throughout the period, plans for the analysis of qualitative data are quite different than quantitative data.

To some extent, the differences in planning for the analysis of qualitative data and quantitative data are the result of the differences in the planning for the sample, measures, and administration of data collection. For quantitative data, the time required for preplanning is often very extensive. For example, with quantitative data the population lists must be assembled in advance for sampling, lists of measures and instruments from prior studies must be identified in advance of developing the instruments for a specific evaluation, and initial drafts of these instruments must be piloted for clarity and minimizing respondent burden. All of these activities lengthen the amount of time that must be incorporated into the plan before

actual data collection. However, these processes often make the analysis more straightforward and less time consuming, not to say that it is without challenges, especially when assumptions about the outputs of the prior data collection processes are violated. One of those assumptions that is frequently violated in the practice of many evaluations is that few data will be missing. The violation of this assumption can produce findings with errors. Indeed, this is so common that most evaluation plans will include procedures to reduce missing data during the data collection and to minimize bias that may result during the analysis process. These plans should be informed by research on how to maximize response rates (Dillman et al., 2014) and minimize bias (Graham, 2009), which is becoming more standard practice in evaluations.

Planning for the analysis of qualitative data often is less straightforward than planning for quantitative data. In recent years, the analysis of qualitative data has become much more systematic and often, in fact, quantifies participant responses. For example, in some cases qualitative evaluators will attempt to provide systematic indications of the pervasiveness of certain experiences among the program participants as well as the strength of the reaction to the experiences that the participants express during the data collection. Making the analysis of qualitative data more systematic often requires iterative processing of the data, sometimes with the assistance of computer programs to count occurrences of key words or themes. This sometimes requires the evaluators who collected the data to comb through the data to extract meanings and compare and contrast them with the data collected from other participants. In other words, the analysis of qualitative data, like the process of collecting qualitative data, can be somewhat open ended, requiring the analyst to exercise judgment and agency, and therefore difficult to fully plan in advance.

Almost by definition, more research questions, more data collected, and more challenges during the data collection process will extend the time needed for data analysis. At this point in planning the evaluation process, it is important to critically think through the tension inherent in evaluation. On one hand, to be influential evaluation findings are needed as soon as possible and certainly in advance of stakeholders' making relevant decisions about the program that the evaluation may be able to guide. On

the other hand, evaluators are responsible for obtaining and systematically analyzing data and providing answers to the research questions that are as accurate as possible. This inherent tension can result in compressing the period of time allocated for data collection and analysis during the planning process. The plan, including the time allocated to carry out all of the processes, should be realistic. If the realistic plan cannot be completed by the time that stakeholders need the information, the plan should be altered so that a less ambitious plan that provides credible and valid information when it is needed can be developed and implemented. Compromise, following open and realistic conversations among stakeholders during planning, is often needed to produce a practical analysis plan.

Communication Plan

Earlier in this chapter, we indicated that the primary goal of evaluation is to influence actions and attitudes. The final component of most evaluation plans is a communication plan, which is essentially a plan to move from findings to influence. Perhaps not surprising to most readers somewhat familiar with evaluations, the main component of the communication plan is the evaluation report. However, the process of developing the report and the nature of the report can take many different paths. In fact, many seasoned evaluators doubt that the time and resources invested in preparing a full report can always be justified. Also, briefings—oral presentations of the evaluation and its findings—are often not considered as important as they actually are. The objective of the evaluators in developing the communications plan is to consider the options for influencing actions and attitudes through reports and stakeholders briefings in a manner that maintains independence of the evaluators and transparency about the findings, evaluation process, data, and data limitations.

The communication plan presents the opportunity for the evaluators to reach agreement with sponsors and other key stakeholders about the release of findings. This is an essential part of the planning process. A first-order question is who has the right to release study information. In some cases, organizations contract for an evaluation and retain the right to decide if and when any public release of any and all information, including findings and recommendations, will occur. When sponsors maintain control of the release of information, other stakeholders and the public may be skeptical about whether the information that is released presents the full and complete findings. To ensure more independence is exercised in the form of a more complete release of information regardless of the nature of the findings, evaluators often retain intellectual property rights to the information generated during the course of conducting the evaluation. Usually, the evaluators agree to provide sponsors and stakeholders with an advance copy of the report or briefing materials they plan to release. Stakeholders and sponsors are given a specific period of time to provide comments and point out any statements they believe are factually incorrect. Then the evaluators present a final version with any revisions they believe

will improve the accuracy of the information to the sponsors and stakeholders, along with a date when the report will be released. In the evaluation plan, it is very important to establish the ownership of the information, including who has the rights to release it publicly, the conditions that must be met in order to release it, and the time periods for any reviews prior to release. It can become very difficult to negotiate these processes and conditions after the evaluation has begun and even more so if the findings are known and not entirely positive. To be sure, evaluators differ on the importance of the release of the findings, but when programs are funded through tax revenue or affect the public, especially those individuals with little power or influence, many evaluation organizations, evaluation theorists, and evaluators believe strongly that the public has a right to access the evaluation findings through its reports and briefing materials.

Reports

The main objective of evaluation reports is to answer the research questions and assist stakeholders in interpreting the findings as well as understanding their implications. In order to accurately interpret the findings, stakeholders will need to understand the data, methods, and processes that produced the findings. For smaller scale evaluations, the final report can be concise and focused, providing the main findings as well as briefly documenting the methods and data. In other cases, especially for longer term and more comprehensive evaluations, multiple reports may be issued from a single evaluation. The reports can be staged over time when findings become available and may influence program decisions, such as changing how a program is implemented. In these cases, the initial reports are often labeled preliminary or interim, and it is helpful to include the study period in the title or subtitle to allow the audience to distinguish them easily from later interim reports and the final report. In other cases, multiple reports can be organized by topic or by the type of evaluation. Choosing a reporting strategy involves trade-offs between timeliness, clarity of focus, and brevity when multiple reports are chosen and a comprehensive and complete report that presents all of the evidence to be brought to bear on the program when a single report is chosen. When multiple reports are chosen, evaluators should be aware that later findings may contradict earlier ones, such as cases in which more immediate outcomes are more positive than later ones. Also, multiple reports generally require more resources, including personnel time for writing, editing, formatting, and publication, than a single report.

In single, comprehensive reports or final reports, some evaluators may attempt to combine all of the evidence they have amassed to provide an overall description of the merit and worth of a program as some of early evaluation theorists advocated (Scriven, 1991). However, current practice appears to lean toward descriptive valuing in which evaluators take a more neutral stance on the overall merit and worth of a program in favor of listing a program's strengths and weaknesses in the findings section. In reports following this approach, evaluators typically present positive and then negative findings separately, without judgment, implications, or

attempting to weight them for an overall conclusion about merit and worth. The objective of this strategy is to allow stakeholders to weigh the evidence and develop their own conclusions. In addition, evaluators are divided on whether recommendations should be offered in the evaluation report. Recommendations for improvement come most directly from the findings of evaluations that include process monitoring, assess implementation fidelity, and assess program design. However, in cases in which needs assessments find significant and important gaps between current program coverage and outcomes and the desired or expected coverage and outcomes or when impacts are negative or neither positive nor negative, evaluators may have garnered insights about how to improve the program during the evaluation. In the cases in which summing up the merit and worth or providing fundamental recommendations are expected from the evaluation, the plan needs to include processes and time to generate and ensure they are thoroughly vetted, because the implications for making overall judgments about the program can affect many individuals and groups with different perspectives on the program.

To facilitate interpretation of the findings, it is necessary for the evaluators to describe the study's research design, sample, measures, data, and data analysis. One of the most important considerations during the report planning process involves the level of detail used to describe the study methods. In some cases, it may be most desirable to provide a full description of the methods, especially when the study may be disseminated broadly and some audiences with stakes in the program may need information about the data to interpret the findings. In these cases, sections on each of the study components will be needed: (a) objectives and research design; (b) sample (for impact evaluations, both the treated and control or comparison samples); (c) data, including source and collection procedures; (d) measures (outcomes, program variables of interest, outputs, and covariates); and (e) data analysis. For other evaluations, especially local, smaller scale, and less technical evaluations, only a few sentences—one or two for each of the five categories just listed—describing the methods may be needed.

Briefings and Interactions

Often stakeholders' attention is more focused on briefings and discussions of the evaluation findings than on reports. Therefore, planning for briefings and organizing interactions can be very important because they can involve substantial amounts of time for the key evaluation personnel, and they can be the locus of the decision making about any actions that will be taken as a result of the evaluation. Parallel to reports, the main purposes of briefings are to describe the findings, aid in their interpretation, and bring their implications to light. To accomplish these objectives, stakeholders will need to have at least a basic understanding of the study's methods and data.

Evaluators can stage and sequence briefings and reports in ways that take advantage of the interactions among stakeholders and between the stakeholders and evaluators. For example, evaluators may conduct a briefing with key stakeholders on preliminary findings to obtain their initial reactions, including insights and additional questions they raise about the program or findings, and use this information to guide the final analyses and development of the report. In other circumstances, the evaluators provide a report to key stakeholders and sponsors and then follow it with a briefing to discuss the report and findings. Briefings can vary in content and format for different groups of stakeholders. In a recent evaluation conducted by one of the authors, both an advisory committee and the leadership and staff of the agency delivering the program were briefed semiannually as new findings about implementation and outcomes were available. The advisory committee received more concise briefings about findings, and they contributed input on the meaning and interpretation of the findings. The intervention leadership and staff received briefings that provided much more detail on the implementation of the intervention and the immediate outcomes such as participant engagement and staff development. The state agency leadership along with the leader of the intervention received briefings in advance of the advisory committee and staff. However, the state agency board received annual briefings that summarized the key findings and suggested improvements in the intervention.

Sometimes reports and briefings are made available simultaneously. Sometimes they are staged to gain buy-in and timeliness. Like the overall evaluation plan, the communication plan should be tailored to fit the evaluation, its political and organizational context, and the actions it was expected to influence.

Project Management Plan

The project management plan serves several purposes. First, it ensures that the skills and time commitments of the evaluation personnel match those needed to undertake the activities described in the previous sections of the plan. Second, the project management plan describes the resources, including equipment and facilities as well as technical support, that will be needed and the sources of these resources during the conduct of the evaluation. Finally, it lays out the key milestones for the evaluation and the date by which the evaluators or stakeholders are expected to accomplish each of them. The latter is usually presented as a study timeline. In cases in which stakeholders are expected to provide data or other resources for the evaluation, such as letters recruiting sites, the timeline may include responsibilities assigned to these stakeholders as well as those assigned to the evaluators.

Personnel

The main purpose of the personnel section is to provide sufficient background on the key members of the evaluation team to demonstrate that they have the skills and time to complete the tasks or, in the case of larger scale evaluations, to oversee the completion of the tasks. The personnel section includes some background on all of the key members of the evaluation team. Sometimes it is difficult to decide who should be included. In general, the overall leaders for the evaluation should be included along with other personnel who are responsible for accomplishing or overseeing the accomplishment of the major tasks included in the timeline. The main ingredients for the background sections are the relevant prior experiences of the study personnel and their responsibilities for the current evaluation. The experience component should at a minimum describe previous evaluations in which the individual has conducted or managed tasks that are similar to those assigned to him or her in the proposed project. It also may list the training, preparation, and experience, including their terminal degree and focus of prior evaluations that have prepared them to undertake and complete their assignments on the proposed evaluation. For individuals with less direct experience, their preparation becomes more salient for establishing their capacity for their assigned tasks.

Resources

Resources include the equipment and facilities as well as the relationships or other things that will support accomplishing the evaluation tasks. For any evaluation plan, this section may include idiosyncratic elements, but a few things are commonly described. When original data collection is called for in the plan, the resources that will be used need to be listed. For example, the software used for Web-based surveys is often listed. When secondary data are to be used, the software, hardware, and actual data may be included in the resources. In cases in which administrative data or a public-use data set is to be used, the computing environment that will allow secure storage and facilitate analysis should be described. Having housed and analyzed the data or data that is quite similar in prior studies may support the adequacy of the computing resources for the current evaluation. Mentioning this can increase confidence that the evaluation team can carry out the plan. In addition to these type of resources, prior working relationships to facilitate data collection, site recruitment, or communication of findings to broad audiences or key stakeholders are important resources that may deserve mention in this section of the plan. Often the resources are backed up with more technical descriptions or letters of agreement in the case of prior relationships. These are often submitted along with the plan in appendices or made available online by the evaluators.

Study Timeline

The study timeline contains the **milestones** or major tasks to be completed during the evaluation. Usually these milestones will align with the major activities described in the plan. Certainly, activities associated with obtaining the study sample, all major steps in each of the data collection activities, the data analysis, often preliminary and final, the reports, and briefings will need to be listed. Each item or milestone has a date for it to be completed and usually an individual or organization responsible for completing it. An example timeline is displayed in [Exhibit 11-C](#).

Exhibit 11-C Example Timeline for an Evaluation Using Mixed Methods for Data Collection (Administrative, Survey, Site Visits, and Documents)

Implementation and Evaluation Tasks	Lead	Year 1											
		8	9	10	11	12	1	2	3	4	5	6	
Program participants identified	Agency		X										
Plan for services finalized	Agency	X											
Meetings with agency and evaluation team	Evaluator												
Evaluation procedures finalized	Evaluator	X											
Develop survey	Evaluator				X								
Develop implementation fidelity items	Evaluator				X								
Develop service report coding rubric	Evaluator					X							
Phase 1 services delivered	Agency												
Site research access agreements	Agency		X										
Phase 2 services delivered	Agency								X				
Phase 3 services delivered	Agency												
Administrative data transferred (baseline from prior year)	Agency												
Administrative data preparation	Evaluator					X							
Update baseline data treatment and control	Evaluator						X						
Check balance and conduct validity checks	Evaluator									X			
Report data transferred	Agency												
Pilot and finalize survey	Evaluator						X						
Test and finalize service report coding	Evaluator												
Semiannual Briefing 1	Evaluator						X						
Evaluation advisory committee meeting	Evaluator												
Code and analyze service report data	Evaluator								X				
Site visits	Evaluator									X			
Administer survey	Evaluator								X				
Preliminary implementation briefing	Evaluator							X					
Clean and analyze survey data	Evaluator										X		
Annual report preparation and finalize	Evaluator												
Update preliminary briefing	Evaluator											X	
Semiannual Briefing 2	Evaluator											X	

Note: Gray shading represents the period in which activities are conducted. X marks the completion month for the task.

Note: Gray shading represents the period in which activities are conducted. X marks the completion month for the task.

Summary

- The main components of an evaluation plan are (a) purpose and scope; (b) data collection, acquisition, and management; (c) data analysis; (d) communication; and (e) project management.
- The purpose and scope list the research questions to be addressed and convey an overview of the study methods, including research design, sample, measures, data, and data analysis.
- The data collection, acquisition, and management section of the plan describes the primary data collection activities, data that will be acquired from secondary sources, and the construction and management of databases.
- For quantitative and qualitative data, the data analysis section lays out the main steps in the analysis of the data that will be undertaken to address each of the research questions.
- The communication section will describe the reports, briefings, and any other form of communication between the evaluators and key stakeholders or the public along with the timing of each.
- The project management section will provide relevant background and qualifications of key personnel, the resources such as equipment and facilities needed to carry out the evaluation, and a timeline listing project milestones with dates when they are expected to be accomplished.
- The level of detail for the plan will vary on the basis of the scale of the program and scope of the evaluation. More comprehensive or larger scale evaluations will require more extensive plans. Less formal evaluations or those of smaller programs may allow less detailed plans. However, plans are essential for the evaluators, the evaluation sponsors, and other key stakeholders.

Key Concepts

Case studies 272

Causal designs 272

Descriptive designs 272

Influence 266

Milestones 287

Primary data 277

Secondary data 277

Standards 271

Critical Thinking/Discussion Questions

1. Discuss the role of influence in program evaluation. What types of influence can an evaluation have? What activities and strategies can be included in the evaluation plan to facilitate the different kinds of influence the findings may have?
2. What elements are essential in writing proper research questions for evaluations?

Application Exercises

For the following questions, locate a final evaluation report. Ideally the evaluation will be of a large-scale social intervention at the state or federal level.

1. What social intervention is being evaluated? What are the research questions addressed in the evaluation? What outcomes were measured?
2. Explain the research design. Was the design descriptive or casual? What makes the research design appropriate for evaluating this particular social intervention?
3. What data were analyzed in the evaluation? How were data acquired? What was the data analysis strategy? Were the data quantitative, qualitative, or both? Why do you think those data sources were chosen? Do you think other data sources should have been included?
4. What were the main findings of this evaluation? After reviewing the report, given what you now know about conducting evaluations, would you recommend that any changes be made to the evaluation plan? If you were to conduct an evaluation in the future on a similar social intervention, what would you do differently?

Chapter 12 The Social and Political Context of Evaluation

The Social Ecology of Evaluations

Multiple Stakeholders

The Range of Stakeholders

Consequences of Multiple Stakeholders

Disseminating Evaluation Results

Evaluation as a Political Process

Political Time and Evaluation Time

Issues of Policy Significance

The Profession of Evaluation

Intellectual Diversity and Its Consequences

The Education of Evaluators

Consequences of Diversity in Origins

Diversity in Working Arrangements

Inside Versus Outside Evaluations

Organizational Roles

The Leadership Role of Elite Evaluation Organizations

Evaluation Standards, Guidelines, and Ethics

Utilization of Evaluation Results

Guidelines for Maximizing Utilization

Epilogue: The Future of Evaluation

Summary

Key Concepts

In the 21st century, evaluation has become ubiquitous and spread throughout the globe. Evaluation is a purposeful activity, designed to improve social conditions by improving policies and programs. This purpose demands that evaluators undertake more than simply applying appropriate research procedures. They must be keenly aware of the social ecology in which the program is situated and that, in the broadest sense of politics, evaluation is a political activity.

As we explain in this chapter, evaluation is characterized by its diversity: Diversity in the intellectual tradition of its practitioners and their training. Diversity in the organizational settings in which they work. Diversity in the scope and methods of their practice and in

their views of how evaluations should influence social programs and policies and the ways in which the influence of evaluations can be enhanced.

Along with the growth of evaluation and its diversity has come a movement toward professionalism. Evaluation is not a profession, but associations have arisen around the world to support evaluators and provide training along with setting guidelines for practice.

Evaluations are a real-world activity. In the end, evaluations should not be judged by the critical acclaim an evaluation receives from peers in the field but the extent to which it leads to the modification of policies, programs, and practices—ones that, in the short or long term, improve the human condition. As long as society continues to believe in the possibility of improving social conditions through the application of knowledge and evidence, we see every reason to believe that the evaluation enterprise will continue to grow.

In *The Evaluation Society*, Peter Dahler-Larsen (2012) begins with a simple declaration: “We live in the age of evaluation” (p. 1). Practicing evaluators, stakeholders, and the public at large must have come to the same conclusion. Evaluation is everywhere—it is one of the growth industries of our time. Dahler-Larsen also points out that with the spread of evaluation, certain tensions have arisen. One such tension, the inconsistent utilization of evaluation, is a problem that has long been recognized and has stimulated innovations in evaluation practice aimed at facilitating its utility. Another issue is that evaluation has become highly diverse. With so many variations and innovations in evaluation practice—qualitative and quantitative, experimental and empowerment, case studies and generalizable samples—and so many diverse individuals with different backgrounds and perspectives conducting evaluations, evaluation sponsors can have difficulty deciding what kind of evaluation they need and may not get what they expected from any particular evaluation team. Another tension between evaluation and the broader society is that evaluation may highlight the complexity of the social problem a program attempts to ameliorate. Rather than offering simple and straightforward solutions that can be enacted by decision makers and administrators, evaluations may point out the limitations of actions taken within the policy silos that exist.

In the 21st century, compared with the late 1970s, when the first edition of this textbook was published, evaluators are more aware of the limitations and challenges posed in conducting evaluations and disseminating the findings. It is evident that simply undertaking well-designed and carefully

conducted evaluations of social programs by itself will not eradicate our human and social problems. But along with the tensions that have arisen as evaluation has become commonplace, the contributions of the evaluation enterprise in moving social intervention in the desired direction should be recognized. There is considerable evidence that the findings of evaluations do often influence policies and programs in beneficial ways, sometimes in the short term and other times in the long term. In this chapter, we take up the complexity surrounding conducting evaluations, the diversity and professionalization of the field, and the continuing challenges and successes in the utilization of evaluations.

The Social Ecology of Evaluations

To conduct successful evaluations, evaluators need to continually assess the complex social ecology of the arena in which they work. Sometimes the impetus and support for an evaluation come from the highest decision-making levels: Congress or a federal agency may mandate evaluations of innovative programs. For example, in 2008, the U.S. Department of Labor contracted for the evaluation of the Adult and Dislocated Worker program authorized by the Workforce Investment Act, which mandated an evaluation (Mathematica, 2008–2017). The evaluation addressed questions about the implementation of the program, its impact on participants' employment and earnings, and the cost-effectiveness of the program. Evaluators conducted the study at 28 randomly chosen local sites in which the outcomes of eligible participants randomly assigned to intensive services or intensive services with training were compared with basic services such as access to local job listings. The short-term findings indicate that the intensive services led to higher earnings, but the addition of training did not increase earnings.

In other cases, the board of a philanthropic foundation may mandate the evaluation of the foundation's major social action programs. For example, the David and Lucile Packard Foundation has established guiding principles for monitoring, evaluation, and learning that specify that the foundation will "track, assess, and learn from our work at multiple levels: individual grants, clusters of grants, strategy, and field. We are selective in our evaluation of individual grants, focusing on those of high cost or high degree of risk, models that could be leveraged, and work with high learning potential for the field" (Packard Foundation, n.d.). At other times, evaluation activities are initiated in response to requests from managers and supervisors of various operating agencies and focus on administrative matters specific to those agencies and stakeholders. At still other times, evaluations are undertaken in response to the concerns of individuals and groups in the community who have a stake in a particular social problem and the planned or current efforts to deal with it.

Whatever the impetus may be, evaluators' work is conducted in a real-world setting of multiple and often conflicting interests. In this regard, two essential features of the context of evaluation must be recognized: the existence of multiple stakeholders and the related fact that evaluation is usually part of a political process.

Multiple Stakeholders

Evaluators usually find that diverse individuals and groups have an interest in their work and its outcomes for a particular program. These stakeholders may hold competing and sometimes combative views about the appropriateness of the evaluation work and about whose interests will be affected by the outcome. To conduct their work effectively and contribute to the resolution of the issues at hand, evaluators must understand their relationships to the stakeholders involved as well as the relationships among stakeholders. The starting point for achieving this understanding is to recognize the range of stakeholders who directly or indirectly can affect the usefulness of evaluation efforts.

The Range of Stakeholders

The existence of a range of stakeholders is as much a fact of life for the lone evaluator situated in a single school, hospital, or social agency as it is for evaluators associated with evaluation groups in large professional research organizations, federal and state agencies, universities, or private foundations. In an abstract sense, every citizen should be concerned with the effectiveness of efforts to improve social conditions and have a stake in the findings of an evaluation. In practice, of course, the stakeholders concerned with any given evaluation effort consist mainly of those with direct and visible interests in the program. Among those, different stakeholders typically have different perspectives on the meaning and importance of an evaluation's findings. These disparate viewpoints are a source of potential conflict not only among stakeholders themselves but also between these individuals and the evaluator. No matter how an evaluation comes out, there are often some for whom the findings are good news and some for whom they are bad news.

To evaluate is to make judgments; to conduct an evaluation is to provide empirical evidence that can be used to inform judgments. The distinction between making judgments and providing information on which judgments can be based is useful and clear in the abstract but often difficult to make in

practice. No matter how well an evaluator's conclusions about the effectiveness of a program are grounded in a rigorous research design and sensitively analyzed data, some stakeholders are likely to perceive those conclusions to be arbitrary or capricious and to react accordingly.

Perhaps the only reliable prediction is that the parties most likely to be attentive to an evaluation, both while it is under way and after a report has been issued, are the evaluation sponsors and the program managers and staff. Of course, these are the groups that usually have the most at stake in the continuation of the program and whose activities are most directly judged by the evaluation. The reactions of the intended beneficiaries of a program may also present a particular challenge or opportunity for an evaluator, depending on their point of view. In many cases, beneficiaries may have the strongest stake in an evaluation's outcome, yet they are often the least prepared to make their voices heard. Target beneficiaries tend to be unorganized and disbursed geographically; often they are grappling with the circumstances that led them to be the intended beneficiaries. Sometimes they are reluctant even to identify themselves. When target beneficiaries do make themselves heard in the course of an evaluation, it is often through organizations that attempt to represent them. For example, homeless persons rarely make themselves heard in the discussion of programs directed at relieving their distressing conditions. But the National Coalition for the Homeless, an organization composed of both persons who themselves are not homeless and current and former homeless individuals, often acts as a spokesperson in policy discussions dealing with homelessness.

Increasingly, evaluators have sought to include intended program beneficiaries as stakeholders in the evaluation. In Australia and New Zealand, evaluators' efforts to include Aboriginal people in the evaluation have progressed from respect for Aboriginal people and cultural competence to turning control and ownership of evaluations over to the Aboriginal people. Participatory evaluation, culturally responsive evaluation, and empowerment evaluation have principles that encourage respect and direct involvement of culturally diverse groups that themselves have been traditionally the subjects of evaluation. Balancing the direct involvement and influence of intended program beneficiaries as

stakeholders and their role as the intended beneficiaries is the subject of ongoing discussion and evolving practices in the field of evaluation (see, e.g., the Center for Culturally Responsive Evaluation and Assessment, at <https://crea.education.illinois.edu>).

Consequences of Multiple Stakeholders

There are two important consequences of the attention of multiple stakeholders to an evaluation. First, evaluators must accept that their contributions as evaluators are but one input into the complex political processes from which decisions and actions eventuate. Second, strains invariably result from the conflicts among the interests of these stakeholders. In part, these strains can be eliminated or minimized by anticipating and planning for them; in part, they come with the turf and must be dealt with on an ad hoc basis or simply accepted and lived with.

The multiplicity of stakeholders generates strains for evaluators in three main ways. First, evaluators are often unsure whose perspective they should take in designing an evaluation. Is the proper perspective that of society, the government agency involved, the program administrators and staff, the program's intended beneficiaries, or one or more of the other stakeholder groups? For some evaluators, especially those who aspire to provide advice for improving programs, the program administrators and staff may be viewed as the primary audience. For evaluators whose projects have been mandated by a legislative body, the primary audience may include the relevant society, whether it is the community, the state, or the nation as a whole.

The issue of which perspective to take in an evaluation should not be understood as one of whose bias to accept. Perspective issues are involved in defining the goals of a program and deciding which stakeholder's concerns should be most closely attended to in relation to those goals. In contrast, bias in an evaluation usually means distorting an evaluation's design or conclusions to favor findings that are in accord with some stakeholder's desires. Every evaluation is undertaken from some set of perspectives, but an ethical evaluator tries to avoid such bias.

In our judgment, the responsibility of the evaluator is not to take one of the many perspectives as the sole legitimate one but, rather, to be clear about the perspective from which a particular evaluation is being undertaken while giving recognition to the other perspectives. In reporting the results of an evaluation, for example, an evaluator can state that the evaluation was conducted from the viewpoint of the program administrators while acknowledging the alternative perspectives of the society as a whole and of the program clients.

In some evaluations, it may be possible to provide several perspectives on a program. Consider, for example, an assistance program for individuals with disabilities who are currently unemployed. From the viewpoint of those individuals, a successful program may be one that provides payment levels sufficient to meet basic consumption needs. From that perspective a program with relatively low levels of payments may be judged as falling short of its aim. But from the perspective of state legislators, for whom the main purpose of the program is to facilitate employment of the clients, the low level of payment may be seen as creating a desirable incentive. By the same token, legislators may view a generous assistance program that might be judged a success from the perspective of the beneficiaries as fostering welfare dependency. With these contrasting views on a central feature of the program, it would be appropriate for the evaluator to be concerned with both kinds of program outcomes: the adequacy of payment levels for basic needs and how payment levels affect employment and independence.

A second way in which the varying interests of stakeholders can generate strain for evaluators concerns the responses to the evaluation findings. Regardless of the perspective used in the evaluation, there is no guarantee that the outcome will be satisfactory to any particular group of stakeholders. Evaluators must realize, for example, that even the sponsors of an evaluation may turn on them when the results do not support the policies and programs they advocate. Although evaluators often anticipate negative reactions from other stakeholder groups, frequently they are unprepared for the responses of the sponsors to findings that are contrary to what these stakeholders expected or desired. Evaluators are in a very difficult position when this occurs. Losing the support of the evaluation sponsors may, for

example, severely constrain the evaluator's ability to conduct other evaluations.

A third source of strain is the misunderstandings that may arise because of difficulties in communicating with different stakeholders. The vocabulary of the evaluation field is no more complicated and esoteric than the vocabularies of the social sciences from which it is derived. But this does not make it understandable and accessible to lay audiences. For instance, the concept of *random* plays an important role in impact assessment. To evaluation researchers, the random assignment of individuals to intervention and control groups means something quite precise, delimited, and valuable. In lay language, however, *random* often calls to mind *haphazard*, *careless*, *aimless*, *casual*, and so on, all with pejorative connotations.

It may be too much to expect an evaluator to master the subtleties of communication to the widely diverse audiences for evaluations. Yet the problem of communication remains an important obstacle to the understanding of evaluation procedures and the utilization of evaluation results. Evaluators are, therefore, well advised to anticipate the communication barriers in relating to stakeholders, a topic we will discuss more fully later in this chapter.

Disseminating Evaluation Results

For evaluation results to be influential, they must be disseminated to and understood by major stakeholders and the general public. For our purposes, *dissemination* refers to the activities through which knowledge about evaluation findings is made available to the relevant audiences.

Dissemination is a critical responsibility of evaluation researchers. An evaluation that is not made accessible to its audiences is destined to be ignored. Accordingly, evaluators must take care in writing their reports and make provision for ensuring that findings are delivered to major stakeholders.

Obviously, evaluation results must be communicated in ways that make them intelligible to the various stakeholder groups. External evaluators generally provide sponsors with technical reports that include detailed and complete descriptions of the evaluation's purpose, design, data collection methods, analysis procedures, results, suggestions for further research, and perhaps recommendations, as well as a discussion of the limitations of the data and analysis. Technical reports usually are read in their entirety only by peers, rarely by the stakeholders who could put the findings to use. Many of these stakeholders simply do not have the time to read voluminous documents and might not be able to understand them, especially the technical details germane to a review by other researchers.

For this reason, every evaluator must learn to be a secondary disseminator. **Secondary dissemination** refers to the communication of results and recommendations that emerge from evaluations in ways that meet the needs of stakeholders (to supplement the **primary dissemination** to sponsors and technical audiences, which in most cases is the technical report). Secondary dissemination may take different forms, including abbreviated summaries of the purpose and study findings, often called executive summaries or research briefs, special reports in more attractive and accessible formats, oral briefings complete with slides, and sometimes even videos.

The objective of secondary dissemination is simple: to provide results in ways that can be comprehended by readers without a grounding in research,

especially interested stakeholders with their different backgrounds and perspectives. Proper preparation of secondary dissemination documents is a part of the craft of evaluation that is garnering more attention during academic training in evaluation and applied research. The important tactic in secondary communication is to find the appropriate style for presenting research findings, using language and form understandable to audiences who are unschooled in the vocabulary and conventions of the evaluation field. *Language* implies a reasonable vocabulary level that is as free as possible from jargon; *form* means that secondary dissemination documents should be succinct, short, and readily comprehensible. Useful advice for this process can be found in Torres, Preskill, and Piontek (2005). In addition, books and courses on data visualization (Evergreen, 2017) and graphical display of data are available that aid evaluators in providing access to data in a non-numerical format. Alternative formats and making the evidence accessible to diverse audiences is important for achieving the main purpose for which evaluation is undertaken, that is, to be used in service of ameliorating social problems.

Evaluation as a Political Process

Throughout this book, we have stressed that evaluation results should be useful in decision making about a program's development and operation. In the earliest phases of program development, evaluations can provide basic data about social problems so that sensitive and appropriate programs can be designed. While prototype programs are being tested, evaluations of pilot demonstrations may provide estimates of the effects to be expected when the program is fully implemented. After programs have been in operation, evaluations can provide evidence about their operational performance and effectiveness. But this is not to say that what is useful in principle will automatically be understood, accepted, and used. At every stage, evaluation is only one ingredient in an inherently political process. And this is as it should be: Program and policy decisions with important social consequences should be determined in a democratic society by political processes.

In some cases, evaluation sponsors may commission an evaluation with the expectation that it will critically influence the decision to continue, modify, or terminate a project. In those cases, the evaluator may be under pressure to produce information quickly, so decisions can be made expeditiously. In other situations, evaluators may complete their assessments of an intervention only to discover that decision makers react slowly to their findings. Even more disconcerting are the occasions when a program is continued, modified, or terminated without regard to an evaluation's relevant and often expensively obtained information.

Although in such circumstances evaluators may feel that their labors have been in vain, they should remember that the results of an evaluation are usually only one input to the decision-making process. The many parties involved in a social program, including sponsors, managers, operators, and clients, often have very high stakes in the program's continuation, and their opinions may count more heavily than the results of the evaluation, no matter how objective it may be.

In any political system that is sensitive to weighing, assessing, and balancing the conflicting claims and interests of different constituencies, the evaluator's role is that of an expert witness, testifying about a program's performance and effectiveness and bolstering that testimony with empirical evidence. A jury of decision-makers and other stakeholders may give such testimony more weight than uninformed opinion or shrewd guessing, but they, not the expert witness, are the ones who must reach a verdict. There are other considerations to be taken into account. To imagine otherwise would be to see evaluators as having the power of veto in the political decision-making process, a power that would strip decision makers of their responsibilities in that regard. In short, the proper role of evaluation is to contribute the best possible knowledge on evaluation issues to the political process and not to attempt to supplant that process.

Political Time and Evaluation Time

There are two additional strains involved in evaluation research compared with academic research that are consequences of the fact that the evaluator is engaged in a political process involving multiple stakeholders. One is the need for evaluations to be relevant and significant in a policy sense, a topic we will take up momentarily; the other is the difference between political time and evaluation time.

Evaluations, especially those directed at assessing program impact, take time. Large-scale impact evaluations that estimate the net effects of major innovative programs may require years to complete. The political and program worlds often move at a much faster pace. Policymakers and project sponsors usually are impatient to know whether a program is achieving its goals, and often their time frame is a matter of months, not years.

For this reason, evaluators frequently encounter pressure to complete their assessments more quickly than the best methods permit, as well as to release preliminary results, perhaps prematurely. At times, evaluators are asked for their impressions of a program's effectiveness, even when they point out that such impressions may prove to be misleading before all the evidence is in. For example, a rigorous evaluation of the Adult and Dislocated Worker program was mandated in the Workforce Investment Act

of 1998. However, the evaluation was not commissioned until 2008, and the findings on the program's effectiveness were not available until 2016, 2 years after it was reauthorized under the Workforce Innovation and Opportunity Act. Thus, reauthorization occurred without the benefit of evidence from the evaluation that the policymakers themselves had commissioned.

In addition, the planning and procedures for initiating evaluations within organizations that sponsor such work often make it difficult to undertake timely studies. In many cases, procedures must be approved at several levels and by a number of key administrators. As a result, it can take considerable time to commission and launch an evaluation, not counting the time it takes to implement and complete it. Although both government and private sector sponsors have tried to develop mechanisms to speed up the planning and procurement processes, these efforts can be hindered by the workings of their bureaucracies, by legal requirements related to contracting, and by the need to establish agreement on the evaluation questions and design.

It is not clear what can be done to reduce the pressure resulting from the different time schedules of evaluators and decision makers. It is important that evaluators anticipate the demands and needs of stakeholders, particularly the evaluation sponsors, and avoid making unrealistic time commitments. Generally, a long-term study should not be undertaken if the information is needed before the evaluation can be completed. One promising innovation that is currently being pursued to increase timeliness and relevance of evaluation findings for making program and policy decisions is the support of evaluation partnerships, or using more official terminology research-practitioner partnerships, between teams of evaluators and local or state education agencies by the Institute of Education Sciences in the U.S. Department of Education. These partnerships support rigorous impact, implementation fidelity, and cost-effectiveness evaluations of educational programs, for example, turning around the lowest performing schools or systematic evaluation of teachers' performance. The support can last up to 5 years and facilitates the exchange of information between evaluators and key local stakeholders regularly throughout the evaluation. For instance, in one such evaluation, the evaluation team provided the

program leadership and staff with information on implementation fidelity, quality, and variability semiannually and within 2 months of the close of a period of data collection.

Issues of Policy Significance

Evaluations, we have stressed, are done with a purpose that is practical and political in nature. In addition to the issues we have already reviewed, the fact that evaluations are ultimately conducted to affect the policy-making process introduces several considerations that distinguish evaluation research from other forms of social science research.

Policy Space and Policy Relevance. The alternatives considered in designing, implementing, and assessing a social program are ordinarily those within the current **policy space**, the set of alternative policies that can garner political support at any given point in time. A difficulty is that policy space keeps changing in response to the efforts of influential figures to gain support from policymakers and from events that refocus policy priorities. For example, in response to the epidemic of school shootings in the United States, some policymakers at the federal and state levels have begun to seriously consider allowing teachers to be armed, an alternative policy proposition that would have been unthinkable just a few years ago.

Because a major purpose of evaluation is to help decision makers form new social policies and to assess the worth of ongoing programs, evaluation research must be sensitive to the various policy issues involved and the limits of policy space. The goals of an evaluation project must resemble those articulated by policymakers in deliberations on the issues of concern. A carefully designed randomized experiment showing that a reduction in certain regressive taxes would lead to an improvement in worker productivity may be irrelevant if decision makers are more concerned with motivating entrepreneurs and attracting potential investments.

For these reasons, responsible impact assessment design must necessarily involve, if at all possible, some contact with relevant decision makers to ascertain their interests in the program being piloted or considered for a demonstration project. A world-wise evaluator will attempt to figure out

what the current and future policy space will allow to be considered. For an innovative project that is not currently under discussion by decision makers, but is being tested because it may become the subject of future discussion, the evaluators and sponsors must rely on their informed forecasts about what changes in policy space are likely. Adjustments to the policy space frequently take the form of moving the line between what is public domain and private domain. Bans on public smoking and requirements to report possible child abuse by health care professionals are examples of renegotiating the line between actions considered to be under the purview of individuals or families and actions restricted by law or policy. Privatizing prisons moves the public operation and administration of correctional facilities to the private sector. Evaluators can consult the proceedings of deliberative bodies (e.g., government committee hearings or legislative debates), interview decision makers' staffs, consult decision makers directly, or review the discourse among the relevant policy community about novel ideas for policy solutions to ongoing social problems. The latter is particularly germane because policy ideas tend to percolate within the community of officials, journalists, academics, and interest groups for some time before entering the space where they become credible alternatives.

Policy Significance. The fact that evaluations are conducted according to the canons of social research may make them more objective than other modes of judging social programs, but they provide only superfluous information unless they address the values of the persons engaged in policy making, program planning, and management. That is, evaluations must have **policy significance**. The weaknesses of evaluations, in this regard, tend to center on how research questions are stated and how findings are interpreted. The issues here involve considerations that go beyond methodology. To maximize the utility of evaluation findings, evaluators must be sensitive to two levels of policy considerations.

First, programs that address problems on the national or state policy agenda, that is, programs that are frequently the subject of legislative hearings or studies or executive policy priorities, require especially close attention from evaluators assessing them. Evaluations of highly visible programs are heavily scrutinized for their methodological rigor and technical proficiency, particularly if they are controversial, which is often the case.

Methodological choices are always matters of judgment and sensitivity to their significance in the policy process. Even when formal economic efficiency analyses are undertaken, the issue remains. For example, the decision to use a participant, program sponsor, or community accounting perspective will be determined largely by policy and stakeholder considerations.

Second, evaluation findings must be assessed according to how far they are generalizable, whether the findings are significant for the policy and for the program, and whether the program clearly fits the need (as expressed by the many factors involved in the policy-making process). An evaluation may produce results that all would agree are statistically significant and generalizable and yet are not sufficiently compelling to be significant for policy, planning, and managerial action. Some of the issues involved in such situations are discussed in detail in [Chapter 9](#) under the rubric of practical significance.

Our hope is that the foregoing observations about the dynamics of conducting evaluations in the context of the real world of social programs and policy sensitize the evaluator to the importance of scouting the terrain when embarking on an evaluation and of staying alert to changes in the social ecology that occur during the evaluation process. Such efforts may be at least as important to the successful conduct of evaluation activities as the technical appropriateness of the procedures employed.

The Profession of Evaluation

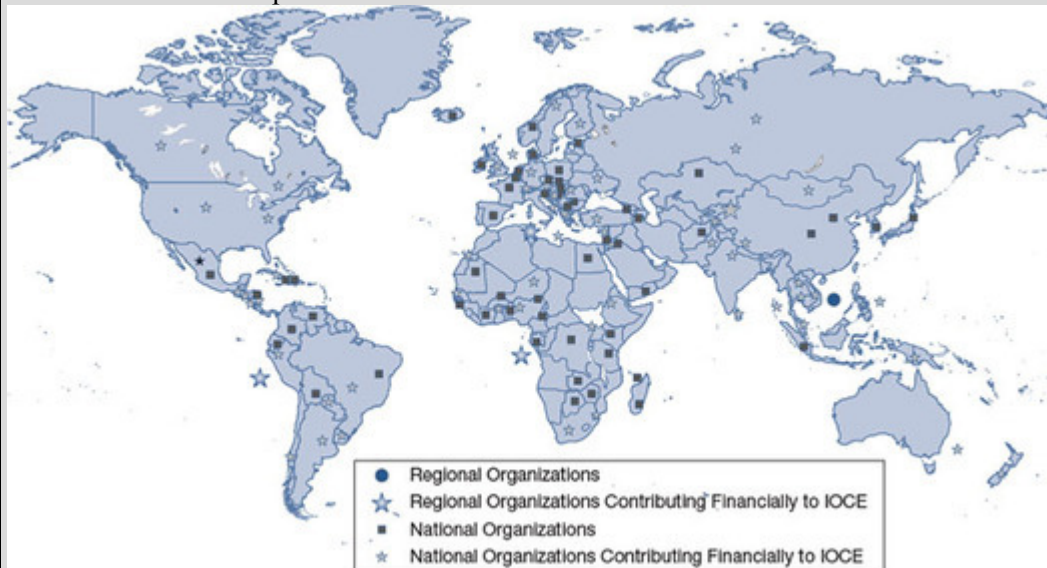
Evaluators work in widely disparate program areas and devote varying amounts of their work time to evaluation activities. Indeed, the labels *evaluator* and *evaluation researcher* conceal the heterogeneity, diversity, and amorphousness of the field. Evaluators are not licensed or certified, so the identification of a person as an evaluator provides no assurance that he or she shares any core knowledge or training with any other person so identified. One of the most noticeable developments in the field of evaluation as it has grown is the large number of national and regional organizations of evaluators. The American Evaluation Association, the major membership organization dedicated to evaluation in the United States, has roughly 7,000 members spread across the United States and 60 other countries. In [Exhibit 12-A](#), we display a map of the locations of 133 evaluation organizations around the globe that have registered with the International Organization for Cooperation in Evaluation.

While growing in numbers and organizations providing networking and development opportunities, evaluation is not a profession by the criteria usually applied to characterize such groups. Much discourse has occurred about evaluator competencies (e.g., King & Stevahn, 2015), but it has yet to be codified into a recognized set of qualifications required for anyone conducting evaluations. It remains accurate to describe evaluators as a collection of individuals sharing a common label, who are not formally organized, and who may have little in common with one another in terms of the range of activities they undertake or their approaches to evaluation, competencies, organizations within which they work, and perspectives. This feature of the evaluation field underlies much of the discussion that follows.

Exhibit 12-A National and Regional Evaluation Organizations

In 2003, representatives of 24 evaluation associations and networks launched the International Organization for Cooperation in Evaluation (IOCE) in Lima, Peru. The global reach of evaluation is evident on the map of national and regional evaluation organizations depicted below. To date, 133 national evaluation organizations have registered with IOCE, in addition to international organizations, multinational organizations, and regional organizations in Africa, Latin America, Australasia, and Europe. The mandate of IOCE is to “contribute to building evaluation leadership and

capacity, especially in developing countries; advance the exchange of evaluation theory and practice worldwide; address international challenges in evaluation; and assist the evaluation profession to take a more global approach to contributing to the identification and solution of world problems.”



Source: Downloaded from the International Organization for Cooperation in Evaluation (<https://www.ioce.net/members>) on July 27, 2018.

Intellectual Diversity and Its Consequences

Evaluation has a richly diverse intellectual heritage. All the major social science disciplines—economics, psychology, sociology, political science, and anthropology—have contributed to the development of the field. And individuals trained in each of these disciplines have made contributions to the concepts and methods of evaluation research. Persons trained in the various professional fields with close ties to the social sciences—public policy, medicine, public health, social work, urban planning, public administration, education, and the like—have also made important contributions and have undertaken significant evaluations. In addition, statistics, biostatistics, econometrics, and psychometrics have contributed important ideas on measurement, causal inference, and analytical techniques.

In the abstract, the diverse roots of the field are one of its strengths: Each disciplinary and professional perspective can add to richness of the options for evaluation practice. At the same time, however, the diverse roots of the field confront evaluators with the need to be general social scientists and lifelong students if they are to keep up, let alone broaden their knowledge base. Clearly, it is impossible for every evaluator to be a scholar in all of the social sciences and to be an expert in every methodological procedure. There is no ready solution to this limitation, but it does mean that evaluators must at times forsake opportunities to undertake work because their knowledge base may be too narrow, or they may have to use a good enough method rather than a more appropriate one with which they are unfamiliar. As the evaluation enterprise has grown, it has also resulted in greater specialization among practicing evaluators around content, method, and approach. This also means that frequently evaluators will need to form teams, not only for the volume of work involved with large-scale evaluation but to ensure that relevant knowledge and skills are represented. Furthermore, it follows that sponsors of evaluations and managers of evaluation staffs must be increasingly knowledgeable about the wide range of evaluation approaches and practices and exercise discretion when selecting contractors and in making work assignments.

In a well-organized profession, a range of opportunities is available for keeping up with the state of the art and expanding one's repertoire of competencies, for example, the peer learning that occurs at regional and national meetings and the didactic courses provided by professional evaluation associations. However, even with the expansion of evaluation associations, it is impossible to know how many of the thousands of individuals undertaking evaluations participate in these organizations and take advantage of the opportunities they provide.

The Education of Evaluators

The diffuse character of the evaluation field is exacerbated by the different ways in which evaluators are educated. Few people working in evaluation have achieved responsible posts and rewards solely by working their way up within dedicated evaluation units. Most evaluators have some sort of formal graduate training either in social science departments or professional schools, but there are very few such programs devoted entirely to evaluation research. In some universities, interdisciplinary programs in evaluation have been set up that include graduate instruction across a number of departments, including Claremont Graduate University and Western Michigan University. In these programs, a graduate student might have the opportunity to take courses in test construction and measurement in a department of psychology, econometrics in a department of economics, survey design and analysis in a department of sociology, policy analysis in a political science department, and evaluation theory and practice courses that could be in almost any social science department or professional studies department, such as public health, education, or social work.

Alternatively, professional schools increasingly offer specializations or tracks that concentrate on evaluation. Schools of education train evaluators for positions in that field, programs in schools of public health train persons who can engage in health service evaluations, and so on. In fact, over time these professional schools have provided much of the formal training evaluators receive that is specifically focused on evaluation theory and practice. However, those programs have their limitations as well. One criticism is that they expose students to a variety of methods and practices but do not provide the conceptual breadth and depth that allows graduates

to develop sensitivity to the social and political context in which evaluations take place. Another is that the courses most relevant to evaluation may be added at the margins of a broader curriculum related to the overarching professional practice (education, public health, etc.), which allows relatively few courses and limited coverage of the relevant concepts and methods. However, variations do occur in some graduate programs, such as public policy programs that offer relevant courses in several social science disciplines, quantitative and qualitative methods, evaluation and policy analysis, and practicum experiences that allow students to conduct actual evaluations for local sponsors and engage with stakeholders.

We see no obvious advantage for one route over the other; each has its advantages and liabilities. Increasingly, it appears that professional schools are becoming the major suppliers of evaluators, at least in part because of the reluctance of graduate social science departments to develop and staff applied research courses and curricula. But these professional schools are far from homogeneous in what they teach, particularly in the approaches to and methods of evaluation they emphasize—thus the continued diversity of the field.

Consequences of Diversity in Origins

The many pathways to becoming an evaluator contribute to the lack of a coherent framework of concepts and methods in the field. That, in turn, accounts at least in part for the differences in the orientations and approaches different evaluators bring to the evaluations they undertake. Whatever the sources, this disciplinary and professional diversity has produced some amount of conflict within the field of evaluation. Evaluators hold divided views on topics ranging from epistemology to the choice of methods and the major goals of evaluation. Some of the major divisions are described briefly below.

Orientations to Primary Stakeholders.

As mentioned earlier in this chapter, evaluators differ about whose perspective should be a priority in an evaluation. A cadre of evaluators trained in the utilization-focused evaluation approach believe that

evaluations should orient toward specific individuals, who have been labeled as the intended users. Usually this means that evaluators should aid program insiders, usually administrators, in understanding and improving their programs. The originator of utilization-focused evaluation, Michael Quinn Patton (2012), offers several tips for selecting the right individuals to whom to orient the evaluation: “Find and involve the right people, those with interest and influence”; “Recruit primary intended users who represent important stakeholder constituencies”; and “Facilitate high-quality interactions among and with primary intended users” (pp. 72–74). This view of evaluation leans heavily toward consultation with program management and gauges the success of the evaluation by the extent to which it informs action to improve programs.

Other evaluators hold that a primary purpose of evaluation should be to help program beneficiaries become empowered. The key steps in empowerment evaluation begin with a community, which could be residents of a village, marginalized individuals who share a common characteristic or orientation, or members of an organization, assessing their needs, identifying their goals, developing a means of reaching their goals such as a program, finding resources and implementing the program, and assessing program implementation and outcomes (Fetterman, Kaftarian, & Wandersman, 2015). This view of evaluation emphasizes the engagement of a community of intended beneficiaries in a collaborative, problem-solving effort characterized by democratic decision making and the pursuit of social justice.

At the other extreme are evaluators who believe that evaluators should mainly serve those stakeholders who fund the evaluation and the broader public good. Indeed, federal agencies or branches of those agencies, such as National Institute of Justice, Institute of Education Sciences, and National Institutes of Health provide support for evaluations that are often conducted by university-based researchers or researchers in large professional research organizations with the purpose of providing evaluations of ongoing or innovative programs that contribute to general knowledge about effective programs that target policy-relevant outcomes.

Our own view is stated earlier in this chapter. We believe that, as much as possible, evaluations ought to be sensitive to the perspectives of all the major stakeholders. Ordinarily, evaluation grants or contracts require that primary attention be given to the evaluation sponsor's definitions of program goals and outcomes. However, such requirements do not exclude other perspectives. We believe that it is the obligation of evaluators to state clearly the aims of each study and to set forth the procedures for garnering and incorporating the perspectives of key stakeholders. When an evaluation has the resources to accommodate several perspectives, multiple perspectives should be used if appropriate.

The Qualitative-Quantitative Division.

Many of those in the evaluation community divide on their methodological preferences and expertise between advocates of qualitative methods and advocates of quantitative methods. However, the relevance of this distinction and the literature that has developed around it have waned significantly in recent years, although not completely. On one side, advocates of qualitative approaches stress the need for intimate knowledge and acquaintance with a program's concrete manifestations in attaining valid knowledge about the program's effects. Qualitative evaluators tend to be oriented toward formative evaluation, that is, making a program work better by feeding information to its managers and sponsors. In addition, they tend to rely on information about the lived experiences of those being served by the program, drawing on ethnographic research traditions. In contrast, quantitatively oriented evaluators often focus on impact assessments and summative evaluations. They focus on measures of program characteristics, processes, and outcomes that allow program effectiveness to be assessed with relative objectivity.

Often the polemics of the past debates have obscured a critical point, namely, that the choice of methods and approaches depends on the evaluation question at hand. We explicitly address this in [Chapter 11](#), noting that when planning an evaluation, evaluators should seek the type of data most suited to the questions to be addressed and the resources, including time, that are available for the evaluation. As we have stressed, qualitative approaches can play critical roles in program design and are important

means of monitoring programs. In contrast, quantitative approaches are generally more appropriate for estimating impact and economic efficiency. In reality, current practice often features mixed methods, combining qualitative data and analysis for certain questions and quantitative data and analysis for others. To make matters more interwoven, sometimes qualitative data are analyzed quantitatively, for example, when counting the number of times a particular program objective is mentioned in interviews. Conversely, some quantitative measures are turned into categorical or qualitative categories for analysis, such as describing students who are below proficiency as a part of an educational reform evaluation.

Thus, it seems fruitless to argue either side of which is the better approach without specifying the evaluation questions to be studied. Fitting the approach to the research purposes is the critical issue; to pit one approach against the other in the abstract results in a pointless dichotomization of the field. Indeed, the use of mixed methods or multiple methods (i.e., surveys, administrative data, focus groups, and interviews) can strengthen the validity of findings if results produced by different methods are congruent or complementary.

Diversity in Working Arrangements

The diversity of the evaluation field is also manifest in the variety of settings and bureaucratic structures in which evaluators work. First, there are two contradictory theses about working arrangements, or what might be called the insider-outsider debate. One position is that evaluators are best off when their positions are as secure and independent as possible from the influence of project management and staff. The other is that sustained contact with the policy and program staff enhances evaluators' work by providing a better understanding of the organization's objectives and activities while inspiring trust in the results of the evaluation.

There are also ambiguities surrounding the role of the evaluator vis-à-vis program staff and groups of stakeholders regardless of whether the evaluator is an organizational insider or outsider. This is a question about the extent to which relations between evaluators and program personnel should resemble the hierarchical structures typical of many organizations or the collegial model that at least ideally characterizes academia. Inevitably, this will follow from the nature of the organizational context within which the evaluator works and the nature of the relationships with the evaluation sponsor and other key stakeholders.

Inside Versus Outside Evaluations

In the past, some experienced evaluators went so far as to state categorically that evaluations should never be undertaken within the organization responsible for administering the program being evaluated, but should always be conducted by an outside team. One reason outsider evaluations may have seemed the desired option is that there were differences in the levels of training and presumed competence of insider and outsider evaluation staffs. These differences have narrowed. Until the 1960s, university-affiliated researchers or research firms conducted the largest share of evaluations, and this arrangement is still prominent today. Since the late 1960s, however, many public service agencies in various program areas have hired researchers and created units that conduct in-house evaluations.

Also, the proportion of evaluations done by smaller private firms and independent consultants has increased markedly. As research positions in both large and small organizations have increased, more persons who are well trained in the social and behavioral sciences have gravitated toward applied research jobs in public agencies and for-profit firms.

Given the increased competence of staff and the visibility and scrutiny of the evaluation enterprise, there is no reason now to favor one organizational arrangement over another. Nevertheless, there remain many critical points during an evaluation when there are opportunities for work to be misdirected and consequently misused irrespective of the type of organization employing the evaluators. The important issue, therefore, is for any evaluation to strike an appropriate balance between technical quality and utility for its purposes, recognizing that those purposes may often be different for internal evaluations than for external ones.

Organizational Roles

Whether evaluators are insiders or outsiders, they need to cultivate clear understandings of their roles with sponsors and program staff. Evaluators' full comprehension of their roles and responsibilities is one major element in the successful conduct of an evaluation effort. Again, the heterogeneity of the field makes it difficult to generalize on the best ways to develop and maintain the appropriate working relations. One common mechanism is to have in place an advisory group, a technical review committee, or one or more external experts to review the evaluation design, implementation, and findings to provide some modicum of oversight for the evaluation process and products. The ways such advisory groups or consultants work depend on whether an inside or an outside evaluation is involved, on the sophistication of both the evaluator and the program staff, and on the relationship with and investment in the reviewers. For example, large-scale evaluations undertaken by federal agencies and major foundations often have advisory groups that meet regularly and assess the quality, quantity, and direction of the work. Some public and private health and welfare organizations with small evaluation units have consultants who provide technical advice to the evaluators or advise agency directors on the appropriateness of the evaluation units' activities, or both.

Sometimes advisory groups and consultants are mere window dressing; we do not condone their use if that is their only function. When members are actively engaged, however, advisory groups can be particularly useful in fostering interdisciplinary evaluation approaches, in adjudicating disputes between program and evaluation staffs, and in defending evaluation findings in the face of concerted attacks by those whose interests are threatened.

The Leadership Role of Elite Evaluation Organizations

A small group of evaluators, numbering perhaps no more than 1,000, constitutes an elite in the field by virtue of the scale of the evaluations they conduct and the size of the organizations for which they work. They are somewhat akin to the physicians who practice in the hospitals of major medical schools. They and their settings are few in number but powerful in establishing the norms for the field. The ways in which they work and the standards of performance in their organizations represent an important version of professionalism that evaluators in other settings may use as role models.

The number of organizations that carry out large-scale or high-profile evaluations with state-of-the-art technical expertise is small, but the size and number of these organizations have grown substantially since the last edition of this book. But in terms of both visibility and evaluation dollars expended, these organizations occupy a strategic position in the field. Most of the large federal evaluation contracts over the years have been awarded to a small group of these firms, such as Abt Associates, Mathematica Policy Research, MDRC, Westat, RAND Corporation, Research Triangle Institute, American Institutes for Research, and the Urban Institute (to name a few). A handful of research units affiliated with universities operate at a similar level: the National Opinion Research Center at the University of Chicago, the Institute for Research on Poverty at the University of Wisconsin, the Joint Center for Poverty Research (University of Chicago and Northwestern University), and the Institute for Social Research at the University of Michigan, for example. In addition, significant numbers of well-trained evaluators work in the evaluation units of federal agencies that contract for and fund evaluation research and a few of the large national foundations.

One of the features of these elite research organizations is a continual concern with the quality of their work. In part, this has come about because of critiques of the efforts of some of these organizations, which in the past were not always conducted at a high standard. But as the surviving

organizations came to dominate the field, at least in terms of large-scale evaluations, and as they found funders increasingly using criteria of technical competence in selecting contractors, their efforts improved markedly from a methodological standpoint. Currently, much of the work conducted by these organizations sets the expectations for high-quality evaluations for the field, particularly for large-scale impact evaluations. Also, the expertise of their staffs has increased, and they now compete for the best-trained researchers interested in evaluation and applied research. Moreover, many have found it to be in their self-interest to encourage staff to publish in professional journals, participate actively in professional organizations, and engage in frontier efforts to improve the state of the art. To the extent that there is a general movement toward professionalism in evaluation, these organizations are its leaders. However, the separation of these organizations from the graduate education that takes place in research universities has limited the exposure of graduate students to large-scale, technically sophisticated evaluation projects during their formal education.

Evaluation Standards, Guidelines, and Ethics

If the evaluation field cannot be characterized as an organized profession in the usual sense, it has nevertheless become increasingly professionalized. One indication of that has been the efforts of relevant professional associations to formulate and publish standards for evaluation work. Two major efforts have been made to provide guidance to evaluators. Under the aegis of the American National Standards Institute, the Joint Committee on Standards for Educational Evaluation (2011) has published *The Program Evaluation Standards: A Guide for Evaluators and Evaluation Users*, now in its third edition. The Joint Committee is made up of representatives from several professional associations, including, among others, the American Evaluation Association, the American Psychological Association, and the American Educational Research Association. Originally set up to deal primarily with educational programs, the Joint Committee expanded its coverage to include all kinds of program evaluation. The *Standards* cover a wide variety of topics ranging from what provisions should appear in evaluation contracts through issues in dealing with human subjects to standards for the analysis of quantitative and qualitative data. Each of the several core standards is accompanied by cases illustrating how the *Standards* can be applied in specific instances.

In another major effort, the American Evaluation Association developed and adopted the *Guiding Principles for Evaluators* in 1994 and subsequently revised them twice, currently under the title of *Evaluator's Ethical Guiding Principles* (American Evaluation Association, 2018). Rather than proclaim standard practices, the *Ethical Guiding Principles* sets out five general principles for evaluators. The principles follow, and the full statements are presented in [Exhibit 12-B](#).

1. *Systematic inquiry*: Evaluators conduct data-based inquiries that are thorough, methodical, and contextually relevant.
2. *Competence*: Evaluators provide skilled professional services to stakeholders.
3. *Integrity and honesty*: Evaluators behave with honesty and transparency in order to ensure the integrity of the evaluation.

4. *Respect for people*: Evaluators honor the dignity, well-being, and self-worth of individuals and acknowledge the influence of culture within and across groups.
5. *Common good and equity*: Evaluators strive to contribute to the common good and advancement of an equitable and just society.

These five principles are elaborated and discussed in the *Ethical Guiding Principles*, although not to the detailed extent found in the Joint Committee's work.

Exhibit 12-B The American Evaluation Association's *Evaluator's Ethical Guiding Principles*

A: Systematic Inquiry: Evaluators conduct data-based inquiries that are thorough, methodical, and contextually relevant.

A1. Adhere to the highest technical standards appropriate to the methods being used while attending to the evaluation's scale and available resources.

A2. Explore with primary stakeholders the limitations and strengths of the core evaluation questions and the approaches that might be used for answering those questions.

A3. Communicate methods and approaches accurately, and in sufficient detail, to allow others to understand, interpret, and critique the work.

A4. Make clear the limitations of the evaluation and its results.

A5. Discuss in contextually appropriate ways the values, assumptions, theories, methods, results, and analyses that significantly affect the evaluator's interpretation of the findings.

A6. Carefully consider the ethical implications of the use of emerging technologies in evaluation practice.

B: Competence: Evaluators provide skilled professional services to stakeholders.

B1. Ensure that the evaluation team possesses the education, abilities, skills, and experiences required to complete the evaluation competently.

B2. When the most ethical option is to proceed with a commission or request outside the boundaries of the evaluation team's professional preparation and competence, clearly communicate any significant limitations to the evaluation that might result. Make every effort to supplement missing or weak competencies directly or through the assistance of others.

B3. Ensure that the evaluation team collectively possesses or seeks out the competencies necessary to work in the cultural context of the evaluation.

B4. Continually undertake relevant education, training or supervised practice to learn new concepts, techniques, skills, and services necessary for competent evaluation practice. Ongoing professional development might include: formal coursework and workshops, self-study, self- or externally-commissioned evaluations of one's own practice, and working with other evaluators to learn and refine evaluative skills and expertise.

C: Integrity: Evaluators behave with honesty and transparency in order to ensure the integrity of the evaluation.

C1. Communicate truthfully and openly with clients and relevant stakeholders concerning all aspects of the evaluation, including its limitations.

C2. Disclose any conflicts of interest (or appearance of a conflict) prior to accepting an evaluation assignment and manage or mitigate any conflicts during the evaluation.

C3. Record and promptly communicate any changes to the originally negotiated evaluation plans, the rationale for those changes, and the potential impacts on the evaluation's scope and results.

C4. Assess and make explicit the stakeholders', clients', and evaluators' values, perspectives, and interests concerning the conduct and outcome of the evaluation.

C5. Accurately and transparently represent evaluation procedures, data, and findings.

C6. Clearly communicate, justify, and address concerns related to procedures or activities that are likely to produce misleading evaluative information or conclusions. Consult colleagues for suggestions on proper ways to proceed if concerns cannot be resolved, and decline the evaluation when necessary.

C7. Disclose all sources of financial support for an evaluation, and the source of the request for the evaluation.

D: Respect for People: Evaluators honor the dignity, well-being, and self-worth of individuals and acknowledge the influence of culture within and across groups.

D1. Strive to gain an understanding of, and treat fairly, the range of perspectives and interests that individuals and groups bring to the evaluation, including those that are not usually included or are oppositional.

D2. Abide by current professional ethics, standards, and regulations (including informed consent, confidentiality, and prevention of harm) pertaining to evaluation participants.

D3. Strive to maximize the benefits and reduce unnecessary risks or harms for groups and individuals associated with the evaluation.

D4. Ensure that those who contribute data and incur risks do so willingly, and that they have knowledge of and opportunity to obtain benefits of the evaluation.

E: Common Good and Equity: Evaluators strive to contribute to the common good and advancement of an equitable and just society.

E1. Recognize and balance the interests of the client, other stakeholders, and the common good while also protecting the integrity of the evaluation.

E2. Identify and make efforts to address the evaluation's potential threats to the common good especially when specific stakeholder interests conflict with the goals of a democratic, equitable, and just society.

E3. Identify and make efforts to address the evaluation's potential risks of exacerbating historic disadvantage or inequity.

E4. Promote transparency and active sharing of data and findings with the goal of equitable access to information in forms that respect people and honor promises of confidentiality.

E5. Mitigate the bias and potential power imbalances that can occur as a result of the evaluation's context. Self-assess one's own privilege and positioning within that context.

Source: Reprinted with permission from American Evaluation Association (2018).

Evaluators should understand that the *Ethical Guiding Principles* do not supersede ethical standards imposed by most human services agencies and universities. These standards, discussed in [Chapter 11](#), involve the protection of human subjects and require review of all research with human subjects, including evaluations, by institutional review boards. Most social research centers and almost all universities have institutional review boards to oversee research involving humans that require research plans be submitted in advance for approval. Almost all such reviews focus on *informed consent*, upholding the principle that research subjects in most cases should be informed about the research in which they are asked to participate and the risks to which they may be exposed, and that they should actively consent to becoming research participants. In addition, most professional associations (e.g., the American Sociological Association, the American Psychological Association) have ethics codes that are applicable as well and may provide useful guides to professional issues such as proper acknowledgment to collaborators, avoiding exploitation of research assistants, and so on.

How to apply such guidelines in pursuing evaluations is both easy and difficult. It is easy in the sense that the guidelines uphold general ethical standards that anyone would follow in all situations but difficult in cases when the demands of the research might appear to conflict with a standard. For example, an evaluator in need of business might be tempted to bid on an evaluation that called for using methods with which he is not familiar, an action that might be in conflict with the second of the *Ethical Guiding Principles*. In another case, an evaluator might worry whether the procedures she intends to use provide sufficient information for participants to understand that there are risks to participation. In such cases, our advice to the evaluator is to consult other experienced evaluators and in any case avoid taking actions that conflict or even appear to conflict with the guidelines.

Utilization of Evaluation Results

In the end, program evaluations must be judged by their utility for supporting responsible decision making that improves social well-being. In one sense, evaluations could themselves be regarded as social interventions; that is, they are expected to help improve social conditions by way of improved policies and programs. It would be fair to judge them on the extent to which they do so. Often, evaluations are expected to improve programs and policies through **direct instrumental use**, which implies modifications to program operations or other actions taken on the basis of the evaluation process or findings. However, although evaluations can inform direct action that improves programs, it has long been recognized that they may also constructively influence the way decision makers think about social problems and the programs that attempt to ameliorate those problems. Carol Weiss, featured for her contributions to program theory in [Chapter 2](#), used the term *enlightenment* to describe the broader and more conceptual use of evaluation. More recently, the terms *use* and *utilization* have been replaced in some evaluation literature by the term **evaluation influence**. Evaluation influence, in one of its original descriptions in that literature, includes all “evaluation consequences that could plausibly lead toward or away from social betterment” (Henry & Mark, 2003, p. 295). Herbert (2014) furthers this theme with the observation that “influence provides a definition and a framework that reflects the full impact of evaluation and a cohesive way to organize theoretical and empirical knowledge of the effect evaluation can have on programs” (p. 394). In their original formulation, Henry and Mark posited that influence can occur at the individual, interpersonal, and collective levels, meaning that evaluation can influence individual attitudes and actions, interpersonal interactions that affect individuals, and collective actions such as putting a social problem on the agenda of a government body or a decision to adopt and fund a social program on the basis of an evaluation of the program (Henry & Mark, 2003; Mark & Henry, 2004). They also noted that evaluations could be used to justify an action that was previously decided upon, potentially a misuse, or to persuade stakeholders to take an action.

Disappointment about the extent of the utilization of evaluations has been a theme in the evaluation literature for decades and remains a concern among active evaluators. In a 2006 survey of 1,140 members of the American Evaluation Association, 68% reported that they considered the nonuse of evaluation results to be a major problem in their personal experiences (Fleischer & Christie, 2009). The responses of this informed sample most likely reflected respondents' perceptions and experiences with the direct instrumental use of evaluation results, which they clearly felt was not strong even in this modern age of evaluation. At the same time, high proportions of these same respondents felt that evaluations did have considerable influence on such organizational aspects as planned change, ability to learn from experience, questioning basic assumptions about practice, and evaluative thinking. These more conceptual uses of evaluation, therefore, may represent the predominant influence of the work of program evaluators despite their aspirations for more direct application of their findings.

We agree that the conceptual utilization of evaluations often provides important inputs into policy and program development, and we do not believe influence of that sort should be viewed as less important. Conceptual utilization may not be as visible to peers or sponsors as direct use, but it can affect the program at issue as well as the community it serves. This impact ranges from sensitizing persons and groups to current and emerging social problems to influencing future program and policy development by contributing to the cumulative results of relevant evaluations. In that regard, it may be more appropriate to think about the conceptual influence of program evaluations in terms of the combined effect of a series of evaluations and related applied research on program and policy conceptions and plans in a particular intervention area rather than attempt to parse out the influence of a single evaluation.

Guidelines for Maximizing Utilization

The research on utilization and the reports of experienced evaluators have identified a number of factors related to the extent to which evaluations are influential, whether direct or conceptual influence. An informative systematic review of the empirical research on such factors was reported recently by Johnson et al. (2009). The results highlighted the importance of stakeholder involvement in facilitating evaluation use. Effective involvement in this context entailed a high level of engagement, interaction, and communication between key stakeholders and evaluators.

The experienced evaluators surveyed by Fleischer and Christie (2009) agreed with that evidence from the utilization research. They rated “involving stakeholders in the evaluation process” as the most important role of the evaluator for facilitating use. Moreover, among the factors believed to most influence evaluation use, large majorities endorsed such items as

- planning for use at the beginning of the evaluation,
- identifying and prioritizing intended uses of the evaluation,
- communicating findings to stakeholders as the evaluation progresses,
- identifying and prioritizing intended users of the evaluation,
- involving stakeholders in the evaluation process,
- developing a communication and reporting plan, and
- interweaving the evaluation into organizational processes and procedures.

Although these factors are relevant to the utilization of program evaluations, it is worth remembering that there are many other relevant and appropriate influences on decisions about programs other than evaluation results. The efforts evaluators make to facilitate use along the lines of the insights described above should be aimed at providing the fullest understanding and appreciation of the implications of the evaluation findings among key decision makers, both the instrumental and conceptual implications. Given the many factors that influence program decisions in the social and political context within which they are made, it is unrealistic

to expect that the evaluation findings will always have clear and direct influence on those decisions.

Epilogue: The Future of Evaluation

There are many reasons to expect program evaluation to be a continuing and even expanding enterprise. Foremost, of course, there is no indication of any decline in the number or severity of social issues and needs that warrant organized intervention in the view of policymakers and concerned citizens alike. The problems presented by the unequal distribution of resources within and between societies, poverty, crime, educational needs and gaps, food insecurity, drug and alcohol abuse, and myriad other such problematic conditions have proved to be obstinate and difficult. And changing conditions are bringing new concerns and adding to the urgency of prior ones, such as climate change, population growth, mass migration, and technologically driven economic dislocation.

Correspondingly, there is no shortage of program and policy initiatives worldwide that attempt to address such problems at the local, regional, national, and international levels. Under these circumstances, questions about the effectiveness of such programs, and how they can be made more effective, would likely be quite sufficient to sustain program evaluation as a source of guidance to decision makers. What the recent decade or two has brought, however, has been more than continuing recognition of the utility of evaluation to assess the performance of ongoing programs. Rather, there has been a rather remarkable rise in respect for the greater potential of programs that have already been evaluated with positive results and that can then be implemented more widely.

Commonly referred to as the *evidence-based practice movement* or sometimes the *evidence-based programs movement*, this development prioritizes the implementation of programs supported by evidence of effectiveness both as new programs and as replacements for existing programs without such evidence. This movement draws on the increased number of impact evaluations conducted in recent years in the behavioral sciences that have produced an accumulation of program models with at least some credible evidence of effectiveness, often coupled with meta-analysis of the associated evaluation studies that documents the scope of the positive effects across multiple studies (see, e.g., Biglan & Ogden, 2008).

The current prevention and intervention literature abounds with articles on evidence-based practice in public health, mental health, criminal justice, substance abuse treatment, education, social work, and other such areas of human service.

Another manifestation of this movement has been the growth of registries that identify the evidence-based programs certified by one authoritative organization or another that has reviewed the evidence supporting programs in the respective focal area. In the United States, one of the best known of these is the Department of Education's What Works Clearinghouse, which lists hundreds of education programs with evidence that meets the standards used to screen candidate programs. Similar registries have been developed for criminal justice programs (CrimeSolutions.gov), substance abuse programs (the National Registry of Effective Programs and Practices), health care and public health (the Cochrane Collaboration), and for numerous more specialized program areas.

Related developments include expanded discussion at the level of state and national oversight bodies about the value of program evaluation for improving the performance of government. In the United States, for example, the 2014 annual report of the Council of Economic Advisers (2014) published with the *Economic Report of the President to Congress* included a full chapter titled "Evaluation as a Tool for Improving Federal Programs" ([Chapter 7](#)). The 2016 revised *Policy on Results* issued by the Treasury Board of Canada included among its objectives that "departments measure and evaluate their performance, using the resulting information to manage and improve programs, policies and services." The European Commission's Directorate-General for Regional Policy (2014) introduced its *Guidance Document on Monitoring and Evaluation* for the programming period from 2014 to 2020 with the observation that "citizens expect to know what has been achieved with public money and want to be sure that we run the best policy. Monitoring and evaluation have a role to play to meet such expectations" (p. 2). The *Queensland Government Program Evaluation Guidelines* issued by the Economics Division of Queensland Treasury and Trade (2014) in Queensland, Australia, affirms that "evaluation is an essential part of the management and delivery of public sector programs. Well-designed evaluations are an essential tool for public

sector agencies to strengthen efficiency of program delivery and to demonstrate the effectiveness of programs in generating outcomes” (p. 2). And these are only a few examples from the many such government documents available.

There is thus little doubt that policymakers and key stakeholders increasingly expect social programs to be able to demonstrate that they are effective, and that the evaluation approaches and methods described in this book are viewed as a means for establishing that accountability. The opportunities for evaluators well versed in those approaches and methods to contribute to this broad, albeit uneven, evidence-oriented movement to improve the effectiveness of social programs can also be expected to expand. We should not underestimate the challenges for evaluators that will come with such expanded roles and responsibilities, but we hope the guidance offered in this book will help prepare those readers who embrace these opportunities to perform capably and effectively.

Summary

- Evaluation has become commonplace in the 21st century, but its expansion has brought tensions with respect to the extent to which its findings are influential, the diversity with which it is practiced, and its ability to provide simple, straightforward programmatic prescriptions to ameliorate complex and resistant social problems.
- Evaluation is directed to a range of stakeholders with varying and sometimes conflicting needs, interests, and perspectives. Evaluators must determine the perspective from which a given evaluation should be conducted, explicitly acknowledge the existence of other perspectives, be prepared for criticism even from the sponsors of the evaluation, and adjust their communication to the requirements of various stakeholders.
- Evaluators must put a high priority on planning for the dissemination of the results of their work. In particular, they need to become “secondary disseminators” who package their findings in ways that are geared to the needs and competencies of a broad range of relevant stakeholders.
- An evaluation is only one ingredient in a political process of balancing interests and coming to decisions concerning social programs and policies. The evaluator’s role is much like that of an expert witness, furnishing the best information possible under the circumstances; it is not the role of judge and jury.
- Two significant strains that result from the political nature of evaluation are (a) the different metrics for political time and evaluation time and (b) the need for evaluations to have policy-making relevance and significance. Evaluators must look beyond considerations of technical excellence and science, mindful of the larger context in which they are working and the purposes being served by the evaluation.

- Evaluation is marked by diversity in disciplinary training, type of schooling, and perspectives on appropriate methods. Although the field's rich diversity is one of its strengths, it also leads to unevenness in competency, lack of consensus on appropriate approaches, and justifiable criticism of the methods used by some evaluators.
- Evaluators are also diverse in their working arrangements. Although there has been considerable debate over whether evaluators should be independent of program staff, there is now little reason to prefer either inside or outside evaluation categorically. What is crucial is that evaluators have a clear understanding of their role in a given situation.
- A small group of elite evaluation organizations and their staffs occupy a strategic position in the field and account for most large-scale evaluations. Their methods and standards of these organizations contribute to the movement toward professionalization of the field.
- With growing professionalization has come a demand for published standards and ethical guidelines for evaluators. Relevant professional organizations have responded by developing guidelines for practice and ethical principles specific to evaluation work.
- Evaluations themselves may be viewed as social programs; that is, evaluations have as a goal to improve social conditions. The findings from evaluations can have direct influence on a program's operation as well as its expansion, adoption, or termination. Evaluations can also serve to enlighten stakeholders and decision makers about the social problem to be addressed by a program, complexities associated with mitigating it, and how a program produces its effects. This broader utilization of evaluations appears to influence policy and program development, as well as social priorities, albeit in ways that are not always easy to trace and often attributable to any single evaluation.
- Evaluation has been a growth industry, and we see no reason for that to abate in the future.

Key Concepts

Direct instrumental use 310

Evaluation influence 310

Policy significance 299

Policy space 299

Primary dissemination 296

Secondary dissemination 296

Critical Thinking/Discussion Questions

1. Discuss the role of stakeholders in evaluations, including the challenges that having multiple stakeholders presents.
2. How should evaluators work with decision makers in terms of conducting evaluations and disseminating the results of an evaluation?
3. Evaluators come from varied educational and professional backgrounds. What are the advantages and disadvantages of this diversity to the field of evaluation as a whole?

Application Exercises

1. Review the American Evaluation Association's *Evaluator's Ethical Guiding Principles*. Explain how you plan to uphold these principles given what you've learned throughout this text.
2. The American Evaluation Association Web site offers a community site where evaluators can share their work in a "community library" (<http://comm.eval.org/browse/communitylibraries>). Choose an entry that addresses an area you are interested in. Prepare a short summary you can share with your classmates.

Glossary

Accessibility:

The extent to which the structural and organizational arrangements facilitate participation in the program.

Accountability:

The responsibility of program staff to provide evidence to stakeholders and sponsors that a program is effective and in conformity with its coverage, service, legal, and fiscal requirements.

Accounting perspectives:

Perspectives underlying decisions on which categories of goods and services to include as costs or benefits in an economic efficiency analysis. Common accounting perspectives are those that take the perspective of program participants, program sponsors and managers, and the community or society in which the program operates.

Administrative data system:

A data system that routinely collects and reports information about the delivery of services to clients and, often, billing, costs, diagnostic and demographic information, and outcome status.

Administrative standards:

Stipulated achievement levels set by program administrators or other responsible parties, for example, intake for 90% of the referrals within 1 month. These levels may be set on the basis of past experience, the performance of comparable programs, or professional judgment.

Articulated program theory:

An explicitly stated version of program theory that is spelled out in some detail as part of a program's documentation and identity or as a result of efforts by the evaluator and stakeholders to formulate the theory.

Assessment of program process:

An evaluative study that answers questions about program operations, implementation, and service delivery. Also known as a process evaluation or an implementation assessment.

Assessment of program theory and design:

An evaluative study that answers questions about the conceptualization, design, and theory of action of a program.

Assignment variable:

In regression discontinuity designs, the quantitative variable that provides values for each unit in the study sample that are used to assign them to intervention or control conditions depending on whether they are above or below a predetermined cut-point value. Also called a *forcing variable* or *cutting-point variable*.

Attrition:

The loss of outcome data measured on individuals or other units assigned to comparison or intervention groups, usually because those individuals cannot be located or refuse to contribute data.

Benefits:

Positive program effects, usually translated into monetary terms in cost-benefit analysis or compared with costs in cost-effectiveness analysis. Benefits may include both direct and indirect effects.

Bias:

As applied to program coverage, the extent to which subgroups of a target population are reached unequally by a program.

Black-box evaluation:

Evaluation of program outcomes without the benefit of an articulated program theory or relevant program process data to provide insight into what is presumed to be causing those outcomes and why.

Case studies:

An approach to evaluations that focuses on a program site or small number of sites in which the program participants and program

context, service delivery and implementation, and outcomes are described.

Causal designs:

Randomized designs, regression discontinuity designs, and all the varieties of comparison group designs that are implemented in evaluations assessing program impact and which provide the estimates of the program effects on the outcomes of interest.

Cluster randomized trial:

A randomized control design for impact evaluation in which aggregate units, such as communities, schools, or clinics, are randomly assigned to intervention and control conditions, with outcomes measured on individuals within those aggregate units.

Comparison group:

A group of individuals or other units not exposed to the intervention, or not yet exposed, and used to estimate the counterfactual outcomes for a group that is exposed to the program. Comparison groups are used in designs in which exposure to the intervention is not controlled as part of the design, as is done in randomized control designs in which the comparison group is typically referred to as a control group.

Confirmation bias:

A cognitive bias in which individuals gather, interpret, or remember information selectively in a way that confirms their preexisting beliefs or hypotheses.

Control group:

A group of individuals or other units assigned in an impact evaluation to the condition that is not provided with access or exposure to the intervention; used to estimate the counterfactual outcomes for a group assigned to receive access to the intervention. Control groups are used in randomized control and regression discontinuity designs in which access to the intervention is controlled as part of the design. Compare with *comparison group*.

Cost analysis:

An itemized description of the full costs of a program, including the value of in-kind contributions, volunteer labor, donated materials, and the like.

Cost-benefit analysis:

An analytical procedure for determining the economic efficiency of a program, expressed as the relationship between costs and outcomes, with the outcomes usually measured in monetary terms.

Cost-effectiveness analysis:

An analytical procedure for determining the economic efficiency of a program, expressed as the cost for achieving one unit of an outcome, often used to compare efficiency across different programs.

Costs:

The monetary value of the inputs, both direct and indirect and both paid or in-kind, required to operate a program.

Counterfactual:

The hypothetical condition in which the individuals (or other relevant units) exposed to a program are at the same time, contrary to fact, not exposed to the program. Can also refer to the counterfactual outcomes: the outcomes that would occur for those individuals in that counterfactual condition.

Covariate:

In the context of impact evaluations, a preintervention baseline descriptive variable characterizing the study sample (intervention and comparison groups) that can be used, among other things, to reduce bias in the intervention effect estimates that is associated with baseline differences between the groups.

Coverage:

The extent to which a program reaches its intended target population.

Demonstration program:

Social intervention projects designed and implemented explicitly to test the value of an innovative program concept.

Descriptive designs:

Evaluation research designs that describe, depending on the purpose of the evaluation, the program participants and program context, service delivery and implementation, and outcomes.

Direct instrumental use:

Actions undertaken to improve program operations or other program modification by decision makers and other stakeholders on the basis of specific ideas and findings from an evaluation.

Discounting:

The treatment of time in valuing costs and benefits of a program in efficiency analyses. It involves adjusting future costs and benefits to their present values and requires choice of a discount rate and time frame.

Distributional effects:

Effects of programs that result in a redistribution of resources among the target population.

Dose-response analysis:

Examination of the relationship between the amount or quality of program exposure and the program outcomes.

Effect size statistic:

A statistical formulation of an estimate of a program effect that expresses its magnitude in a standardized form comparable across outcome measures using different units or scales. Two of the most commonly used effect size statistics are the *standardized mean difference* and the *odds ratio*.

Effective sample size:

The operative sample size in statistical power analysis for multilevel impact evaluation designs with assignment at the cluster level and outcomes measured on units within those clusters. Similarity among individuals within clusters makes their outcome data partially redundant (statistically dependent). The effective sample size, which is smaller than the actual total sample size, adjusts for that redundancy.

Effectiveness evaluation:

An impact evaluation of a program that is implemented and operated as routine practice at typical scale and serving a typical target population, that is, not set up as a research or demonstration program. Compare with *efficacy evaluation*.

Efficacy evaluation:

An impact evaluation of a program that is implemented and operated as a research or demonstration program, typically for purposes of determining the ability of the program to produce the intended effects under relatively favorable conditions. The program may be administered and/or evaluated by the program developer. Also known as a proof-of-concept study. Compare with *effectiveness evaluation*.

Efficiency assessment:

An evaluative study that answers questions about program costs in comparison to either the monetary value of its benefits or its effectiveness for bringing about changes in the social conditions it addresses. See also *cost-benefit analysis* and *cost-effectiveness analysis*.

Empowerment evaluation:

A participatory or collaborative evaluation in which the evaluator's role includes consultation and facilitation directed toward the development of the capabilities of the participating stakeholders to conduct evaluations on their own, to use the results effectively for advocacy and change, and to have influence on a program that affects their lives.

Evaluability assessment:

Negotiation and investigation undertaken jointly by the evaluator, the evaluation sponsor, and possibly other stakeholders to determine whether a program meets the preconditions for evaluation and, if so, how the evaluation should be designed to ensure maximum utility.

Evaluation influence:

The direct or indirect effect of evaluation on the attitudes and actions of stakeholders and decision makers.

Evaluation questions:

Questions developed by the evaluator, evaluation sponsor, and/or other stakeholders that define the issues the evaluation will investigate.

Evaluation questions should be stated in terms that can be answered using methods available to the evaluator and in a way useful to stakeholders.

Evaluation sponsor:

The person, group, or organization that requests or requires an evaluation and provides the resources to conduct it.

Ex ante efficiency analysis:

An efficiency (cost-benefit or cost-effectiveness) analysis undertaken before program implementation, usually as part of program planning, to estimate net effects in relation to costs.

Ex post efficiency analysis:

An efficiency (cost-benefit or cost-effectiveness) analysis undertaken after a program's effects are known.

External validity:

The extent to which an estimate of a program effect derived from a subset of the program's target population also characterizes the effect for the full target population, that is, generalizes to that population.

Focus group:

A small panel of persons selected for their knowledge or perspective on a topic of interest that is convened to discuss the topic with the assistance of a facilitator. The discussion is used to identify important themes or to construct descriptive summaries of views and experiences on the focal topic.

Formative evaluation:

An evaluative study undertaken to furnish information that will guide program improvement.

Fundamental problem of causal inference:

The outcome when exposed to the causal factor and the outcome when not exposed cannot both be observed at the same time for the same individuals, but it is the difference between those outcomes that defines the causal effect. See also *potential outcomes* and *program effect*.

Impact:

See *program effect*.

Impact evaluation:

An evaluative study that answers questions about program impact on the outcomes or social conditions the program is intended to ameliorate; that is, the change in outcomes attributable to the program. Also known as an impact assessment.

Impact theory:

A causal theory describing cause-and-effect sequences in which certain program activities are the instigating causes and certain changes in the individuals or other units exposed to the program are the effects they are expected to produce.

Implementation failure:

A situation in which a program does not adequately perform the activities and functions specified in the program design that are assumed to be necessary for bringing about the intended benefits.

Implementation fidelity:

The extent to which the program adheres to the program theory and design and usually includes measures of the amount of service received by the participants and the quality with which those services are delivered.

Implicit program theory:

Assumptions and expectations about how a program brings about its intended effects that are inherent in a program's services and practices but have not been fully articulated and recorded.

Incidence:

The number of new cases of a particular problem, condition, or event that arise in a specified area during a specified period of time.
Compare *prevalence*.

Independent evaluation:

An evaluation in which the evaluator has the primary responsibility for developing the evaluation plan, conducting the evaluation, and disseminating the results but has no role in developing or operating the program.

Influence:

A defining characteristic of evaluations is that they are conducted to influence attitudes and actions. Evaluations can influence individual attitudes or actions, interpersonal behaviors, or collective actions.

Intent-to-treat (ITT) effects:

The program effect estimates that result from a comparison of the outcomes of the intervention and control groups as they were originally assigned to those conditions. The intervention group in intent-to-treat comparisons thus includes those assigned to the intervention who did not receive the intervention along with those who did. Similarly, the control group includes those assigned to the control who did receive the intervention along with those who did not.
Compare with *treatment-on-the-treated* effects.

Interfering event:

In the context of time series designs, an event that occurs at about the same time as the initiation of the intervention with potential to affect the outcome and thus bias the estimate of the intervention effect on that outcome.

Internal rate of return:

The calculated value for the discount rate necessary for total discounted program benefits to equal total discounted program costs.
See *discounting*.

Internal validity:

The extent to which the direction and magnitude of an estimate of a causal effect on an outcome, such as a program effect, are an accurate representation of the unknowable true effect. Internal validity for program effects is presumed to be high when complete outcome data are available for individuals exposed to the program and counterfactual outcomes are estimated with little or no bias.

Interrupted time series:

In impact evaluation, a set of repeated measures of the outcome that begins before the initiation of an intervention and continues afterward, with the intervention thus intruding into the time series in a way that may allow its effects on the outcome to be estimated.

Intervention group:

A group of individuals or other units that are exposed to an intervention and whose outcome measures are compared with those of a *comparison* or *control* group. See also *program group*.

Key informants:

Persons whose personal or professional position gives them a knowledgeable perspective on the nature and scope of a social problem or a target population and whose views are obtained via interviews or surveys.

Matching:

A procedure for constructing a comparison group by selecting individuals or other relevant units not exposed to the program that are identical on specified characteristics to those in an intervention group except for receipt of the intervention.

Maturation:

Natural changes in the individuals or units involved in an impact evaluation of a sort expected to influence the outcomes of interest, for example, the increased abilities of children as they age.

Mediator variable:

In an impact assessment, a proximal outcome that changes as a result of exposure to the program and then, in turn, influences a more distal

outcome. The mediator is thus an intervening variable that provides a link in the causal sequence through which the program brings about change in the distal outcome.

Meta-analysis:

An analysis of effect size statistics derived from the quantitative results of multiple intervention studies for the purpose of summarizing and comparing the findings of that set of studies.

Milestones:

Major tasks and the dates when they are expected to be accomplished throughout the course of an evaluation.

Minimum detectable effect size (MDES):

The smallest effect size determined by some appropriate assessment to have practical significance in the context of a particular program and a given outcome; specified in the form of a standardized statistical effect size. Impact evaluations should be designed to have adequate statistical power for detecting at a statistically significant level any program effect as large as or larger than the MDES.

Moderator variable:

In an impact assessment, a variable, such as gender or age, that characterizes subgroups of the target population for which program effects may differ.

Monitoring and evaluation:

The practice of ongoing collection and reporting of data on program activities, products, and outcomes along with resource utilization and staffing for managing the program combined with outcome or impact evaluation at appropriate points in the life cycle of the program.

Needs assessment:

An evaluative study that answers questions about the social conditions a program is intended to address, the appropriate target population, and the nature of the need for the program.

Negative side effect:

An unintended adverse effect of a program intended to produce beneficial effects; may accompany beneficial effects.

Net benefits:

The total discounted benefits minus the total discounted costs. Also called net rate of return.

Odds ratio:

An effect size statistic that expresses the odds of a successful outcome for the intervention group relative to that of the control group.

Opportunity costs:

The monetary value of opportunities forgone because of involvement of some sort in an intervention program.

Organizational plan:

Assumptions and expectations about what the program must do to bring about the interactions between the target population and the program that will produce the intended changes in social conditions. The program's organizational plan is articulated from the perspective of program management and encompasses both the functions and activities the program is expected to perform and the human, financial, and physical resources required for that performance.

Outcome:

The state of the target population or the social conditions a program is expected to change.

Outcome change:

The difference between outcome levels at different points in time. See also *outcome level*.

Outcome level:

The status of an outcome at some point in time. See also *outcome*.

Outcome monitoring:

Periodic measurement and reporting of indicators of the status of the social condition or outcomes for program participants the program is

accountable for improving.

Participatory or collaborative evaluation:

An evaluation organized as a team project in which the evaluator and representatives of one or more stakeholder groups work collaboratively in developing the evaluation plan, conducting the evaluation, and disseminating or using the results.

Performance criterion:

The standard against which an indicator of program performance is compared so that the program performance can be evaluated.

Policy significance:

The significance of an evaluation's findings for policy and program decisions or assumptions (as opposed to their statistical significance).

Policy space:

The set of policy alternatives that are within the bounds of acceptability to policymakers at a given point in time.

Population at risk:

The individuals or units in a specified area with characteristics indicating that they have a significant probability of having or developing a particular condition or experience. Compare *population in need*.

Population in need:

The individuals or units in a specified area that currently manifest a particular problematic condition or experience. Compare *population at risk*.

Potential outcomes:

An outcome status that would become manifest under certain conditions. The potential outcomes framework for causal inference defines the effect of a known cause as the difference between the potential outcome that would appear with exposure to the cause (e.g., a program) and the potential outcome that would appear without exposure to that cause (e.g., no exposure to the program).

Prevalence:

The total number of existing cases with a particular condition in a specified area at a specified time. Compare *incidence*.

Primary data:

Data collected during the course of an evaluation specifically to address the research questions set forth for the evaluation.

Primary dissemination:

Dissemination of the detailed findings of an evaluation to sponsors and technical audiences.

Probability sample:

A sample from a population in which every member of that population has a known, nonzero chance of being selected for the sample. This means that selection into the sample is done randomly so that it is a matter of chance without any systematic bias in the selection process.

Process evaluation:

Examination of what a program is, the activities undertaken, who receives services or other benefits, the consistency with which it is implemented in terms of its design and across sites, and other such aspects of the nature and operation of the program.

Process monitoring:

Process evaluation that is done repeatedly over time with a focus on selected key performance indicators.

Process theory:

The combination of the program's *organizational plan* and its *service utilization plan* into an overall description of the assumptions and expectations about how the program is supposed to operate.

Program effect:

That portion of an outcome change that can be attributed uniquely to a program, that is, with the influence of other sources controlled or removed; also termed the program's impact. See also *outcome change*.

Program evaluation:

The application of social research methods to systematically investigate the effectiveness of social intervention programs in ways that are adapted to their political and organizational environments and are designed to inform social action to improve social conditions.

Program group:

A group of individuals or other units that receive a program and whose outcome measures are compared with those of a comparison or control group. See also *intervention group*.

Program impact:

See *program effect*.

Program monitoring:

The periodic measurement or documentation of aspects of program performance that are indicative of whether the program is functioning as intended or according to an appropriate standard.

Propensity score:

A score that estimates the probability that an individual or other relevant unit is in the intervention group rather than the comparison group that can be used in various ways to try to reduce selection bias. Propensity scores are constructed from preintervention baseline covariates in a separate analysis before the estimation of the intervention effect.

Quantitative assignment variable:

A variable with at least four unique values on each side of a cutoff that assigns units to an intervention such that all units on one side of a cutoff receive access to the intervention and no units on the other side of the cutoff receive access.

Quasi-experiment:

An impact evaluation design in which intervention and comparison groups are formed by a procedure other than random assignment.

Random assignment:

Assignment of the units in the study sample for an impact evaluation to intervention and control groups on the basis of chance so that every unit in that sample has a known, nonzero probability of being assigned to each group. Also called *randomization*.

Randomized control design:

An impact evaluation design in which intervention and control groups are formed by random assignment and compared on outcome measures to estimate the effects of the intervention. Also called randomized designs, randomized control trials (RCTs), and randomized experiments. See *random assignment*.

Rate:

The occurrence or existence of a particular condition expressed as a proportion of units in the relevant population (e.g., deaths per 1,000 adults).

Regression discontinuity design:

An impact evaluation design in which intervention and control groups are formed on the basis of their scores on a quantitative assignment variable; the units on one side of a cut-point value on the assignment variable are assigned to the intervention condition, and those on the other side are assigned to the control condition. Also known as a cutting-point design. See *assignment variable*.

Regression to the mean:

A phenomenon that can bias estimates of intervention effects in which, with repeated measurement, an extreme value will tend to be followed by a more typical less extreme one (i.e., regress to the mean of the series). Regression to the mean can also occur when individuals are chosen on the basis of a measured variable from the tail of the distribution of scores for the sample from which they are drawn; the value of a subsequent measure will tend to be less extreme, regressing to the mean of the parent distribution.

Reliability:

The extent to which a measure produces the same results when used repeatedly to measure something that has not changed.

Sample survey:

A survey administered to a sample of units in the population. The results are extrapolated to the entire population of interest by statistical projections.

Sampling error:

The chance component introduced into an outcome measure because of the luck of the draw that produced the particular sample from the universe of samples that could have been selected to provide that outcome data. The primary determinant of sampling error is the size of the sample; larger samples are less likely to differ from one another than smaller samples.

Sampling frame:

A list of the units in a population from which a sample is drawn, typically used for a probability sample.

Secondary data:

Data collected before an evaluation, often for administrative purposes, which can be analyzed to address the research questions set forth for the evaluation.

Secondary dissemination:

Dissemination of summarized often simplified findings of evaluations to audiences composed of stakeholders.

Secondary effects:

Effects of a program that impose costs on persons or groups who are not the intended beneficiaries of the program.

Secular trends:

Natural trends in a population of individuals or other units that can bias intervention effect estimates, especially in time series designs. Examples of secular trends are demographic changes in the population resident in a geographical area, changes in economic conditions, increases or decreases in the prevalence of a health condition, and the like.

Selection bias:

Systematic misestimation of program effects that results from uncontrolled differences between a group of individuals exposed to the program and a comparison group not exposed that would result in differences in the outcome even if neither group was exposed to the program. See *counterfactual*.

Sensitivity:

The extent to which the values on a measure change when there is a change or difference in the thing being measured.

Service utilization plan:

Assumptions and expectations about how the target population will make initial contact with the program and be engaged with it through the completion of the intended services. In its simplest form, a service utilization plan describes the sequence of events through which the intended clients are expected to interact with the intended services.

Shadow prices:

Imputed or estimated costs of goods and services not valued accurately in the marketplace. Shadow prices also are used when market prices are inappropriate because of regulation or externalities. Also known as *accounting prices*.

Snowball sampling:

A nonprobability sampling method in which each person who participates in an initial sample is asked to suggest additional people appropriate for the sample, who are then asked to make further suggestions. This process continues until no new names of appropriate persons are suggested.

Social indicator:

A series of periodic measurements designed to track the course of a social condition over time.

Social research methods:

Social science techniques of systematic observation, measurement, sampling, research design, data collection, and data analysis for

producing valid, reliable, and precise characterizations of social behavior.

Stakeholders:

Individuals, groups, or organizations with a significant interest in how well a program functions, for example, those with decision-making authority over the program, funders and sponsors, administrators and personnel, and clients or intended beneficiaries.

Standardized mean difference:

A standardized effect size statistic that expresses the difference between the means for the intervention and control groups on an outcome variable in standard deviation units.

Standards:

The level of performance a program is expected to achieve to be judged adequate.

Statistical power:

The probability that an observed program effect will be statistically significant when, in fact, it represents a real effect. If a real effect is not found to be statistically significant, a Type II error results. Thus, statistical power is one minus the probability of a Type II error. See also *Type II error*.

Summative evaluation:

Evaluative activities undertaken to render a summary judgment on certain critical aspects of a program's performance, for example, to determine if specific goals and objectives were met.

Target population:

The population of units (individuals, families, communities, etc.) to which a program intervention is directed. All such units within the area served by a program constitute its target population.

Targeted program:

A program with a target population defined around specific characteristics or eligibility requirements that constrain who can

receive services. Those constraints may relate to current conditions (e.g., low income, diagnosed mental illness) or indicated risk for an adverse outcome the program aims to prevent. Compare *universal program*.

Theory failure:

A situation in which a program is implemented as planned, but it does not produce the expected effects on the outcomes or the social benefits intended.

Treatment-on-the-treated (TOT) effects:

The program effect estimates that result from a comparison of the outcomes of the units that received the intervention and those that did not receive the intervention, irrespective of the condition to which they were originally assigned. The intervention group in treatment-on-the-treated comparisons thus includes only those who actually received the intervention, and the control group includes only those who in fact did not receive the intervention. Compare with *intent-to-treat (ITT) effects*.

Type I error:

A statistical conclusion error in which an effect estimate is found to be statistically significant when, in fact, there was no actual effect on the respective outcome variable.

Type II error:

A statistical conclusion error in which an effect estimate is not found to be statistically significant when, in fact, there was an effect on the respective outcome variable. See *minimal detectable effect size*, *statistical power*.

Universal program:

A program with a target population that is defined with few or no constraints (e.g., programs in public parks open to all who wish to participate, afterschool programs that accept any child in the school district parents wish to enroll). Compare *targeted program*.

Validity:

When used to describe a measure, the extent to which it actually measures what it is intended to measure.

References

A

- Ajzen, I., & Fishbein, M. (1980). Understanding attitudes and predicting social behavior. Englewood Cliffs, NJ: Prentice Hall.
- Altschuld, J. W. (2010). Needs assessment: Phase II: Collecting data. Thousand Oaks, CA: Sage.
- Altschuld, J. W., & Kumar, D. D. (2010). Needs assessment: An overview. Thousand Oaks, CA: Sage.
- American Bar Association. (2011). ABA standards for criminal justice: Treatment of prisoners. Chicago: American Bar Association, Criminal Justice Standards Committee.
- American Evaluation Association. (2018). Evaluator's ethical guiding principles. Retrieved from <https://www.eval.org/d/do/4220>
- Andersen, S. C., Humlum, M. K., & Nandrup, A. B. (2016). Increasing instruction time in school does increase learning. Proceedings of the National Academy of Sciences, 113(27), 7481–7484.
- Arbogast, J. W., Moore-Schiltz, L., Jarvis, W. J., Harpster-Hagen, A., Hughes, J., & Parker, A. (2016). Impact of a comprehensive workplace hand hygiene program on employer health care insurance claims and costs, absenteeism, and employee perceptions and practices. Journal of Occupational and Environmental Medicine, 58(6), 231–240.
- Arrieta, A., Woods, J. R., Qiao, N., & Jay, S. J. (2014). Cost-benefit analysis of home blood pressure monitoring in hypertension diagnosis

and treatment: An insurer perspective. *Hypertension*, 64(4), 891–896.
doi:10.1161/HYPERTENSIONAHA.114.03780

B

- Bastian, K. C., Henry, G. T., Pan, Y., & Lys, D. (2016). Teacher candidate performance assessments: Local scoring and implications for teacher preparation program improvement. *Teaching and Teacher Education*, 59, 1–12. doi:10.1016/j.tate.2016.05.008
- Berk, R. A., & Sherman, L. W. (1988). Police responses to family violence incidents: An analysis of an experimental design with incomplete randomization. *Journal of the American Statistical Association*, 83(401), 70. doi:10.2307/2288920
- Bernal, N., Carpio, M. A., & Klein, T. J. (2017). The effects of access to health insurance: Evidence from a regression discontinuity design in Peru. *Journal of Public Economics*, 154, 122–136. doi:10.1016/j.jpubeco.2017.08.008
- Bickman, L. (1987). The functions of program theory. *New Directions for Program Evaluation*, 1987(33), 5–18. doi:10.1002/ev.1443
- Biglan, A., & Ogden, T. (2008). The evolution of evidence-based practices. *European Journal of Behavior Analysis*, 9(1), 81–95.
- Blamey, A. A., Macmillan, F., Fitzsimons, C. F., Shaw, R., & Mutrie, N. (2013). Using programme theory to strengthen research protocol and intervention design within an RCT of a walking intervention. *Evaluation*, 19(1), 5–23. doi:10.1177/1356389012470681
- Bloor, M., Leyland, A., Barnard, M., & McKeganey, N. (1991). Estimating hidden populations: A new method of calculating the prevalence of drug-injecting and non-injecting female street prostitution. *Addiction*, 86(11), 1477–1483. doi:10.1111/j.1360-0443.1991.tb01733.x

Boardman, A. E., Greenberg, D. H., Vining, A. R., & Weimer, D. L. (2018). Cost-benefit analysis: Concepts and practice (5th ed.). New York: Cambridge University Press.

Borenstein, M., Hedges, L. V., Higgins, J.P.T., & Rothstein, H. R. (2009). Introduction to meta-analysis. Hoboken, NJ: John Wiley.

Brewer, M. B. (1996, July 1). Donald T. Campbell social psychologist and scholar 1916-1996. Observer. Retrieved from <https://www.psychologicalscience.org/observer/donald-t-campbell-social-psychologist-and-scholar-1916-1996>

Brown, K. M., Stewart, N., & D'Amico, E. (2014). North Carolina Regional Leadership Academies final 2013 activity report. Retrieved July 3, 2018, from http://cerenc.org/wp-content/uploads/2011/10/RLA-Year-2-Report_03_06_2014.pdf

Burch, P., & Heinrich, C. J. (2016). Mixed methods for policy research and program evaluation. Thousand Oaks, CA: Sage.

C

- Caldwell, M. F., Vitacco, M., & Van Rybroek, G. J. (2006). Are violent delinquents worth treating? A cost-benefit analysis. *Journal of Research in Crime and Delinquency*, 43(2), 148–168.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
- Chen, H. (1990). Issues in constructing program theory. *New Directions for Program Evaluation*, 1990(47), 7–18. doi:10.1002/ev.1551
- Chen, H., & Rossi, P. H. (1980). The multi-goal, theory-driven approach to evaluation: A model linking basic and applied social science. *Social Forces*, 59(1), 106. doi:10.2307/2577835
- Chow, M. Y., Li, M., & Quine, S. (2010). Client satisfaction and unmet needs assessment. *Asia Pacific Journal of Public Health*, 24(2), 406–414. doi:10.1177/1010539510384843
- Christie, C. A., & Alkin, M. C. (2003). The user-oriented evaluator's role in formulating a program theory: Using a theory-driven approach. *American Journal of Evaluation*, 24(3), 373–385.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Cohen, J., & Dupas, P. (2010). Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment. *Quarterly Journal of Economics*, 125(1), 1–45.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Chicago: Rand McNally.

Cordray, D. S. (1993). Strengthening causal interpretations of nonexperimental data: The role of meta-analysis. *New Directions for Program Evaluation*, 1993(60), 59–96. doi:10.1002/ev.1661

Correia, A., & Melbin, A. (2005). *Transitional housing services for victims of domestic violence: A report from the Housing Committee of the National Task Force to End Sexual and Domestic Violence*. Washington, DC: National Task Force to End Sexual and Domestic Violence Against Women.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., . . . Weiner, S. S. (1980). *Toward reform of program evaluation*. San Francisco, CA: Jossey-Bass.

D

Dahler-Larsen, P. (2012). *The evaluation society*. Palo Alto, CA: Stanford University Press.

Dahler-Larsen, P. (2017). Theory-based evaluation meets ambiguity. *American Journal of Evaluation*, 39(1), 6–23.
doi:10.1177/1098214017716325

Das, J., Chowdhury, A., Hussam, R., & Banerjee, A. V. (2016). The impact of training informal health care providers in India: A randomized controlled trial. *Science*, 354(6308), aaf7384.
doi:10.1126/science.aaf7384

Davies, R. (2013). *Planning evaluability assessments: A synthesis of literature with recommendations (Working Paper 40)*. London: Department for International Development.

Deke, J., Cook, T., Dragoset, L., Reardon, S., Titunik, R., Todd, P., & Wadell, G. (2015). *Preview of regression discontinuity design standards*. Washington, DC: What Works Clearinghouse.

Diaz, J. J., & Handa, S. (2006). An assessment of propensity score matching as a nonexperimental impact estimator. *Journal of Human Resources*, 41(2), 319–345. doi:10.3368/jhr.xli.2.319

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method (4th ed.)*. Hoboken, NJ: John Wiley.

Donaldson, S. I. (2007). Program theory–driven evaluation science: Strategies and applications. New York: Taylor & Francis.

Dong, N., & Maynard, R. A. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and sample size requirements for experimental and quasi-experimental designs. *Journal of Research on Educational Effectiveness*, 6(1), 24–67.

E

Economics Division, Queensland Treasury and Trade. (2014). Queensland government program evaluation guidelines. Retrieved from <https://s3.treasury.qld.gov.au/files/qld-government-program-evaluation-guidelines.pdf>

European Commission, Directorate-General for Regional Policy. (2014). Guidance document on monitoring and evaluation. Retrieved from http://ec.europa.eu/regional_policy/sources/docoffic/2014/working/wd_2014_en.pdf

Evergreen, S.D.H. (2017). Effective data visualization: The right chart for the right data. Thousand Oaks, CA: Sage.

Executive Office of the President, Council of Economic Advisers. (2014). Economic report of the president together with the annual report of the Council of Economic Advisers. Retrieved from <https://www.gpo.gov/fdsys/pkg/ERP-2014/pdf/ERP-2014.pdf>

F

Federal Judicial Center, Advisory Committee on Experimentation in the Law. (1981). *Experimentation in the law: Report of the Federal Judicial Center Advisory Committee on Experimentation in the Law*. Washington, DC: Author.

Fetterman, D. M., Kaftarian, S. J., & Wandersman, A. (2015). *Empowerment evaluation: Knowledge and tools for self-assessment and accountability* (2nd ed.). Thousand Oaks, CA: Sage.

Fleischer, D. N., & Christie, C. A. (2009). Evaluation use: Results from a survey of U.S. American Evaluation Association members. *American Journal of Evaluation*, 30(2), 158–175.

Fowler, F. J. (2014). *Survey research methods*. London: Sage Ltd.

Fuller, S. C., Roy, M., Belskaya, O., & Leyland, E. (2015). *Local education agency Race to the Top expenditures final analysis of expenditure patterns and related outcomes*. Greensboro: Consortium for Educational Research and Evaluation – North Carolina.

G

Glasser, W. (1975). *Reality therapy: A new approach to psychiatry*. New York: HarperCollins.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60(1), 549–576.
doi:10.1146/annurev.psych.58.110405.085530

H

- Harris, R., Van Dyke, E. R., Ton, T.G.N., Nass, C. A., & Buchwald, D. (2016). Assessing needs for cancer education and support in American Indian and Alaska Native communities in the northwestern United States. *Health Promotion Practice, 17*(6), 891–898.
- Hatry, H. P. (2014). *Transforming performance measurement for the 21st century*. Washington, DC: The Urban Institute.
- Hatry, H. P. (2015). Using agency records. In K. E. Newcomer, H. P. Hatry, & J. S. Wholey (Eds.), *Handbook of practical program evaluation* (4th ed.). San Francisco, CA: Jossey-Bass.
- Hawkins, A. J., Blanchard, V. L., Baldwin, S. A., & Fawcett, E. B. (2008). Does marriage and relationship education work? A meta-analytic study. *Journal of Consulting and Clinical Psychology, 76*(5), 723–734. doi:10.1037/a0012584
- Heinrich, C. J., & Brill, R. (2015). Stopped in the name of the law: Administrative burden and its implications for cash transfer program effectiveness. *World Development, 72*, 277–295. doi:10.1016/j.worlddev.2015.03.015
- Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments? *Psychological Methods, 1*(2), 154–169. doi:10.1037//1082-989x.1.2.154
- Henry, G. T. (1990). *Practical sampling* (Vol. 21, Applied Social Research Methods Series). Newbury Park, CA: Sage.

- Henry, G. T., Campbell, S. L., Thompson, C. L., Patriarca, L. A., Luterbach, K. J., Lys, D. B., & Covington, V. M. (2013). The predictive validity of measures of teacher candidate programs and performance. *Journal of Teacher Education*, 64(5), 439–453. doi:10.1177/0022487113496431
- Henry, G. T., & Mark, M. M. (2003). Beyond use: Understanding evaluations influence on attitudes and actions. *American Journal of Evaluation*, 24(3), 293–314. doi:10.1016/s1098-2140(03)00056-0
- Henry, G. T., Rickman, D. K., Ponder, B. D., Henderson, L. W., Mashburn, A. J., & Gordon, C. S. (2005). *The Georgia Early Childhood Study: 2001-2004: Final report*. Atlanta: Georgia State University, School of Policy Studies.
- Herbert, J. L. (2014). Researching evaluation influence: A review of the literature. *Evaluation Review*, 38(5), 388–419.
- Higher Education Access Tracker. (2017). *Evidencing the impact of outreach participation on student outcomes: Outreach, KS4 attainment and HE progression (Report No. HEAT008)*. Canterbury, UK: Author.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945–960.
- Holvoet, N., van Esbroeck, D., Inberg, L., Popelier, L., Peeters, B., & Verhofstadt, E. (2018). To evaluate or not: Evaluability study of 40 interventions of Belgian development cooperation. *Evaluation and Program Planning*, 67, 189–199. doi:10.1016/j.evalprogplan.2017.12.005
- House, E. R., & Howe, K. R. (1999). *Values in evaluation and social research*. Thousand Oaks, CA: Sage.

- House, L. D., Tevendale, H. D., & Martinez-Garcia, G. (2017).
Implementing evidence-based teen pregnancy-prevention interventions in
a community-wide initiative: Building capacity and reaching youth.
Journal of Adolescent Health, 60(3S), S18–S23.
doi:10.1016/j.jadohealth.2016.08.013
- Houtenville, A. J., & Brucker, D. L. (2014). Participation in safety-net
programs and the utilization of employment services among working-age
persons with disabilities. *Journal of Disability Policy Studies*, 25(2), 91–
105. doi:10.1177/1044207312474308
- Hueftle, S. (2014). Community needs assessment survey St. Paul, NE.
Holdrege, NE: South Central Economic Development District.

I

Institute of Education Sciences. (2015). Request for applications: Education research grants (CFDA No. 84.305A). Retrieved from https://ies.ed.gov/funding/pdf/2019_84305A.pdf

J

- Jacob, R., Somers, M., Zhu, P., & Bloom, H. (2016). The validity of the comparative interrupted time series design for evaluating the effect of school-level interventions. *Evaluation Review*, 40(3), 167–198.
doi:10.1177/0193841x16663414
- Johnson, K., Greeneid, L. O., Toal, S. A., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation*, 30(3), 377–410.
- Johnston, W. R., Harbatkin, E., Herman, R., Migacheva, K., & Henry, G. T. (2018). Measuring fidelity of implementation of a statewide school turnaround intervention: The development of a valid and reliable measure. Washington, DC: Society for Research on Educational Effectiveness.
- Joint Committee on Standards for Educational Evaluation. (2011). The program evaluation standards: A guide for evaluators and evaluation users (3rd ed.). Kalamazoo, MI: Author.

K

- Keehley, P., Medlin, S., Longmire, L., & MacBride, S. A. (1997). Benchmarking for best practices in the public sector: Achieving performance breakthroughs in federal, state, and local agencies. San Francisco, CA: Jossey-Bass.
- Kessler, C., Lamb, M., Stehman, C., & Frazer, D. (2013). Key informant interview report: A summary of key informant interviews and focus groups in Milwaukee County. Retrieved from <https://www.froedtert.com/upload/docs/giving/community-benefit/2012-milwaukee-county-chna-key-informant-interview-report.pdf>
- Kettner, P. M., Moroney, R., & Martin, L. L. (2017). Designing and managing programs: An effectiveness-based approach. Thousand Oaks, CA: Sage.
- Kho, A., Henry, G., Zimmer, R., & Pham, L. (2018). How has iZone teacher recruitment affected the performance of other schools? Nashville: Tennessee Education Research Alliance.
- King, J. A., & Stevahn, L. (2015). Competencies for program evaluators in light of adaptive action: What? So what? Now what? *New Directions for Evaluation*, 145, 21–37.
- Kish, L. (1995). *Survey sampling*. New York: John Wiley.
- Knowlton, L. W., & Phillips, C. C. (2013). *The logic model guidebook: Better strategies for great results*. Thousand Oaks, CA: Sage.

Kuehn, D., Anderson, T., Lerman, R., Eyster, L., Barnow, B., & Briggs, A. (2017). A cost-benefit analysis of Accelerating Opportunity. Washington, DC: Urban Institute.

L

- Lemons, C. J., Fuchs, D., Gilbert, J. K., & Fuchs, L. S. (2014). Evidence-based practices in a changing world: Reconsidering the counterfactual in education research. *Educational Researcher*, 43(5), 242–252.
- Levin, H. M., Belfield, C., Hollands, F., Bowden, A. B., Cheng, H., Shand, R., . . . Hanisch-Cerda, B. (2012). Cost-effectiveness analysis of interventions that improve high school completion. In *Cost-effectiveness analysis of Talent Search* (Chap. 3). New York: Center for Benefit-Cost Studies of Education, Teachers College, Columbia University.
- Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications* (2nd ed.). Thousand Oaks, CA: Sage.
- Leviton, L. C., Khan, L. K., Rog, D., Dawkins, N., & Cotton, D. (2010). Evaluability assessment to improve public health policies, programs, and practices. *Annual Review of Public Health*, 31, 213–233.
- Li, H., Graham, D. J., & Majumdar, A. (2013). The impacts of speed cameras on road accidents: An application of propensity score matching methods. *Accident Analysis & Prevention*, 60, 148–157. doi:10.1016/j.aap.2013.08.003
- Lim, S. S., Dandona, L., Hoisington, J. A., James, S. L., Hogan, M. C., & Gakidou, E. (2010). India's Janani Suraksha Yojana, a conditional cash transfer programme to increase births in health facilities: An impact evaluation. *Lancet*, 375(9730), 2009–2023. doi:10.1016/s0140-6736(10)60744-1
- Lindo, J., & Packham, A. (2015). How much can expanding access to long-acting reversible contraceptives reduce teen birth rates? (Working Paper

No. 21275). Cambridge, MA: National Bureau of Economic Research.
doi:10.3386/w21275

Lippman, L., Anderson Moore, K., Guzman, L., Ryberg, R., McIntosh, H., Ramos, M., . . . Kuhfeld, M. (2014). *Flourishing children: Defining and testing indicators of positive development*. New York: Springer.

Lipsey, M. W. (1993). Theory as method: Small theories of treatments. *New Directions for Program Evaluation*, 1993(57), 5–38. doi:10.1002/ev.1637

Lipsey, M. W., & Pollard, J. A. (1989). Driving toward theory in program evaluation: More models to choose from. *Evaluation and Program Planning*, 12(4), 317–328. doi:10.1016/0149-7189(89)90048-7

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, 48(12), 1181–1209. doi:10.1037//0003-066x.48.12.1181

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Liu, X. S. (2014). *Statistical power analysis for the social and behavioral sciences: Basic and advanced techniques*. New York: Routledge.

M

- Mack, V. (2015). New Orleans kids, working parents, and poverty. New Orleans, LA: The Data Center. Retrieved August 22, 2018, from https://www.datacenterresearch.org/reports_analysis/new-orleans-kids-working-parents-and-poverty/
- MacKinnon, D. P. (2008). Introduction to statistical mediation analysis. New York: Lawrence Erlbaum.
- Mark, M. M., & Henry, G. T. (2004). The mechanisms and outcomes of evaluation influence. *Evaluation*, 10(1), 35–57. doi:10.1177/1356389004042326
- Mark, M. M., Henry, G. T., & Julnes, G. (2000). Evaluation: An integrated framework for understanding, guiding, and improving policies and programs. San Francisco, CA: Jossey-Bass.
- Martin, L. L., & Kettner, P. M. (1996). Measuring the performance of human service programs. Thousand Oaks, CA: Sage.
- Mathematica. (2008-2017). Workforce Investment Act Adult and Dislocated Worker Programs gold standard evaluation. Retrieved from <https://www.mathematica-mpr.com/our-publications-and-findings/projects/wia-gold-standard-evaluation>
- Mavranouzouli, I., Megnin-Viggars, O., Cheema, N., Howlin, P., Baron-Cohen, S., & Pilling, S. (2013). The cost-effectiveness of supported employment for adults with autism in the United Kingdom. *Autism*, 18(8), 975–984. doi:10.1177/1362361313505720

- McDavid, J. C., Huse, I., & Hawthorn, L. R. (2013). Program evaluation and performance measurement: An introduction to practice (2nd ed.). Thousand Oaks, CA: Sage.
- Mielke, K. W., & Swinehart, J. W. (1976). Evaluation of the Feeling Good television series: Summary. New York: Children's Television Workshop.
- Miller, G., & Holstein, J. A. (Eds.). (1993). Constructionist controversies: Issues in social problems theory. New York: Aldine de Gruyter.
- Mishan, E. J., & Quah, E. (2007). Cost-benefit analysis (5th ed.). London: Routledge.
- Morgan, R. E., & Kena, G. (2016). Criminal victimization, 2016. Retrieved from <https://www.bjs.gov/content/pub/pdf/cv16.pdf>
- Morgan, S. L., & Winship, C. (2014). Counterfactuals and causal inference: Methods and principles for social research (2nd ed.). Cambridge, UK: Cambridge University Press.
- Morrel-Samuels, S., Rupp, L. A., Eisman, A. B., Miller, A. L., Stoddard, S. A., Franzen, S. P., . . . Zimmerman, M. A. (2018). Measuring the implementation of Youth Empowerment Solutions. *Health Promotion Practice*, 19(4), 581–589. doi:10.1177/1524839917736511
- Murphy, K. R., Myers, B., & Wolach, A. (2014). Statistical power analysis: A simple and general model for traditional and modern hypothesis tests (4th ed.). New York: Routledge.
- Mustafa, N. (2018). Cost-effectiveness analysis: Educational interventions that reduce the incidence of HIV/AIDS infection in Kenyan teenagers.

International Journal of Educational Development, 62, 264–269.

N

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). The Belmont report: Ethical principles and guidelines for the protection of human subjects of research. Washington, DC: U.S. Department of Health, Education, and Welfare, Office of the Secretary.

Neuman, S. B., & Knapczyk, J. J. (2018). Reaching families where they are: Examining an innovative book distribution program. Urban Education. Retrieved from <http://journals.sagepub.com/eprint/hWhkIUpHuzUc5DTCCcSj7/full>

Nyborg, K. (2014). Project evaluation with democratic decision-making: What does cost–benefit analysis really measure? Ecological Economics, 106, 124–131.

P

Packard Foundation. (n.d.). Guiding principles and practices for monitoring, evaluation and learning. Retrieved from <https://www.packard.org/wp-content/uploads/2017/05/Monitoring-Learning-and-Evaluation-Guiding-Principles.pdf>

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

Patton, M. Q. (2012). Essentials of utilization-focused evaluation. Thousand Oaks, CA: Sage.

Pawson, R. (2013). The science of evaluation: A realist manifesto. Thousand Oaks, CA: Sage.

Petros, S. G. (2011). Use of a mixed methods approach to investigate the support needs of older caregivers to family members affected by HIV and AIDS in South Africa. *Journal of Mixed Methods Research*, 6(4), 275–293. doi:10.1177/1558689811425915

Petrosino, A., Turpin-Petrosino, C., Hollis-Peel, M. E., & Lavenberg, J. G. (2013). “Scared Straight” and other juvenile awareness programs for preventing juvenile delinquency. *Cochrane Database of Systematic Reviews*, 2013(4), CD002796. doi:10.4073/csr.2013.5

Peyton, D. J., & Scicchitano, M. (2017). Devil is in the details: Using logic models to investigate program process. *Evaluation and Program Planning*, 65, 156–162. doi:10.1016/j.evalprogplan.2017.08.012

Puma, M., Bell, S., Cook, R., & Heid, C. (2010). Head Start Impact Study final report. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families.

R

- Riccoboni, S. T., & Darracq, M. A. (2018). Does the U stand for useless? The urine drug screen and emergency department psychiatric patients. *Journal of Emergency Medicine*, 54(4), 500–506.
- Ross, H. L., Campbell, D. T., & Glass, G. V. (1970). Determining the social effects of a legal reform: The British “Breathalyser” crackdown of 1967. *American Behavioral Scientist*, 13(4), 493–509.
- Rossi, P. H. (1990). *Down and out in America: The origins of homelessness*. Chicago: University of Chicago Press.
- Rossi, P. H. (1994). Troubling families. *American Behavioral Scientist*, 37(3), 342–395. doi:10.1177/0002764294037003003
- Rossi, P. H. (1997). Program outcomes: Conceptual and measurement issues. In E. J. Mullen & J. Magnabosco (Eds.), *Outcome and measurement in the human services: Cross-cutting issues and methods*. Washington DC: National Association of Social Workers.
- Rossi, P. H. (2003, November). The “iron law of evaluation” reconsidered. Paper presented at the AAPAM Research Conference, Washington, DC.
- Rossi, P. H., Fisher, G. A., & Willis, G. (1986). *The condition of the homeless of Chicago: A report based on surveys conducted in 1985 and 1986*. Amherst: University of Massachusetts Social and Demographic Research Institute.
- Rossmo, K., & Routledge, R. (1990). Estimating the size of criminal populations. *Journal of Quantitative Criminology*, 6(3), 293–314.

Rossow, I., Storvoll, E. E., Baklien, B., & Pape, H. (2011). Effect and process evaluation of a Norwegian community prevention project targeting alcohol use and related harm. *Contemporary Drug Problems*, 38(3), 441–466. doi:10.1177/009145091103800306

Rutman, L. (1980). *Planning useful evaluations: Evaluability assessment*. Beverly Hills, CA: Sage.

S

- Saunders, R. P. (2016). *Implementation monitoring and process evaluation*. Thousand Oaks, CA: Sage.
- Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage.
- Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, 103(484), 1334–1344.
doi:10.1198/016214508000000733
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth Cengage Learning.
- Shadish, W., & Luellen, J. (2004). Donald Campbell: The accidental evaluator. *Evaluation Roots*, 81–87. doi:10.4135/9781412984157.n4
- Skidmore, F. (1985). Overview of the Seattle-Denver Income Maintenance Experiment: Final report. In L. H. Aiken & B. H. Kehrner (Eds.), *Evaluation studies review annual*. Beverly Hills, CA: Sage.
- Smit, F., Toet, J., & van der Heijden, P. (1997). Estimating the number of opiate users in Rotterdam using statistical models for incomplete count data. In G. Hay, N. McKeganey, & E. Birks (Eds.), *Final report EMCDDA project methodological pilot study of local level prevalence estimates*. Glasgow, UK: Centre for Drug Misuse Research, University of Glasgow/EMCDDA.

Smith, G.C.S., & Pell, J. P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials. *BMJ*, 327(7429), 1459–1461. doi:10.1136/bmj.327.7429.1459

Smith, M. F. (1989). *Evaluability assessment: A practical approach*. New York: Springer.

Snyder, A., Marton, J., McLaren, S., Feng, B., & Zhou, M. (2017). Do high fidelity wraparound services for youth with serious emotional disturbances save money in the long-term? *Journal of Mental Health Policy and Economics*, 20(4), 167–175.

Spector, M., & Kitsuse, J. I. (1977). *Constructing social problems*. Hawthorne, NY: Aldine de Gruyter.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1), 1–21. doi:10.1214/09-sts313

T

Thomas, R. M., Jr. (1996, May 12). Donald T. Campbell, master of many disciplines, dies at 79. The New York Times. Retrieved from <http://jsaw.lib.lehigh.edu/campbell/obituary.htm>

Thompson, C. L., Henry, G., & Preston, C. (2016). School turnaround through scaffolded craftsmanship. *Teachers College Record*, 118(13), 1–26.

Thurston, W., & Potvin, L. (2003). Evaluability assessment: A tool for incorporating evaluation in social change programmes. *Evaluation*, 9(4), 453–469.

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.

Treasury Board of Canada. (2016). Policy on results. Retrieved from <http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=31300>

U

U.S. General Accounting Office. (1986). Teen-age pregnancy: 500,000 births a year but few tested programs (GAO/PEMD-86-16BR). Washington, DC: Author.

U.S. General Accounting Office. (1990). Case study evaluations. Washington, DC: Author.

V

VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction*. New York: Oxford University Press.

VanderWeele, T. J. (2016). Mediation analysis: A practitioner's guide. *Annual Review of Public Health*, 37, 17–32.

Van der Wees, P. J., Zaslavsky, A. M., & Ayanian, J. Z. (2013). Improvements in health status after Massachusetts health care reform. *Milbank Quarterly*, 91(4), 663–689.

Volpp, K. G., Troxel, A. B., Pauly, M. V., Glick, H. A., Puig, A., Asch, D. A., . . . Weiner, J. (2009). A randomized, controlled trial of financial incentives for smoking cessation. *New England Journal of Medicine*, 360(5), 699–709. doi:0.1056/NEJMsa0806819

W

- Watkins, R., Meiers, M. W., & Visser, Y. L. (2012). A guide to assessing needs: Essential tools for collecting information, making decisions, and achieving development results. Washington, DC: World Bank.
- Weber, V., Bloom, F., Pierdon, S., & Wood, C. (2008). Employing the electronic health record to improve diabetes care: A multifaceted intervention in an integrated delivery system. *Journal of General Internal Medicine*, 23(4), 379–382. doi:10.1007/s11606-007-0439-2
- Weiss, C. H. (1972). *Evaluation research: Methods of assessing program effectiveness*. Englewood Cliffs, NJ: Prentice Hall.
- Weiss, C. H. (1979). The many meanings of research utilization. *Public Administration Review*, 39(5), 426. doi:10.2307/3109916
- Weiss, C. H. (1997). Theory-based evaluation: Past, present, and future. *New Directions for Evaluation*, 1997(76), 41–55. doi:10.1002/ev.1086
- West, S. L., & O’Neal, K. K. (2004). Project D.A.R.E. outcome effectiveness revisited. *American Journal of Public Health*, 94(6), 1027–1029. doi:10.2105/ajph.94.6.1027
- Wholey, J. S. (1979). *Evaluation—Promise and performance*. Washington, DC: Urban Institute.
- Wholey, J. S. (1987). Evaluability assessment: Developing program theory. *New Directions for Program Evaluation*, 1987(33), 77–92. doi:10.1002/ev.1447

Wholey, J. S. (2015). Exploratory evaluation. In J. S. Wholey, H. P. Hatry, & K. Newcomer (Eds.), *Handbook of practical program evaluation* (4th ed.). San Francisco, CA: Jossey-Bass.

Wilson, K. J., Brown, H. S., & Bastida, E. (2015). Cost-effectiveness of a community-based weight control intervention targeting a low-socioeconomic-status Mexican-origin population. *Health Promotion Practice*, 16(1), 101–108.

Wilson, S. J., Lipsey, M. W., & Derzon, J. H. (2003). The effects of school-based intervention programs on aggressive behavior: A meta-analysis. *Journal of Consulting and Clinical Psychology*, 71(1), 136–149.

Winfrey, L. T., Esbensen, F., & Osgood, D. W. (1996). Evaluating a school-based gang-prevention program. *Evaluation Review*, 20(2), 181–203. doi:10.1177/0193841x9602000204

Wright, E. R., Ruel, E., Fuoco, M. J., Trouteaud, A., Sanchez, T., LaBoy, A., . . . Hartinger-Saunders, R. (2016). 2015 Atlanta Youth Count and Needs Assessment. Atlanta: Georgia State University.

Z

- Zarling, A., Lawrence, E., & Marchman, J. (2015). A randomized controlled trial of acceptance and commitment therapy for aggressive behavior. *Journal of Consulting and Clinical Psychology, 83*(1), 199–212. doi:10.1037/a0037946
- Zimmer, R., Henry, G. T., & Kho, A. (2017). The effects of school turnaround in Tennessee's achievement school district and innovation zones. *Educational Evaluation and Policy Analysis, 39*(4), 670–696. doi:10.3102/0162373717705729
- Zimmer, R. W., & Guarino, C. M. (2013). Is there empirical evidence that charter schools “push out” low-performing students? *Educational Evaluation and Policy Analysis, 35*(4), 461–480. doi:10.3102/0162373713498465
- Zimmer, R., Gill, B., Booker, K., Lavertu, S., Sass, T., & Witte, J. (2009). Charter schools in eight states: Effects on achievement, attainment, integration, and competition (Report No. MG-869). Santa Monica, CA: RAND.

Author Index

Ajzen, I., [67](#)
Alkin, M. C., [65](#)
Altschuld, J. W., [17](#), [34](#), [55](#)
Ambron, S. R., [2](#)
American Bar Association, [271](#)
American Evaluation Association, [307](#), [308–309](#)e
Arbogast, J. W., [3](#)
Asch, D. A., [201](#)

Baldwin, S. A., [218](#)
Barnard, M., [42](#)
Bastian, K. C., [128](#)
Belfield, C., [261](#)
Bell, S., [268](#)
Belskaya, O., [269](#)
Berk, R. A., [194](#)
Bernal, N., [206](#)
Bickman, L., [65](#)
Biglan, A., [313](#)
Blanchard, V. L., [218](#)
Bloom, H., [180](#)
Bloor, M., [42](#)
Boardman, A. E., [239](#)
Borenstein, M., [232](#)
Bowden, A. B., [261](#)
Brill, R., [270](#), [273](#)
Burch, P., [278](#)

Campbell, D. T., [165](#), [166e](#), [228](#)
Campbell, S. L., [127](#)
Carpio, M. A., [206](#)
Chen, H. -T., [65](#), [68](#), [82](#)
Cheng, H., [261](#)

Christian, L. M., [43](#), [280](#), [282](#)
Christie, C. A., [65](#), [311](#)
Clark, M. H., [182](#)
Cohen, J., [3](#), [220](#)
Cook, R., [268](#)
Cook, T., [204](#)
Cook, T. D., [165](#)
Cordray, D. S., [85](#)
Correia, A., [271](#)
Cotton, D., [62e](#)
Council of Economic Advisers, [313](#)
Covington, V. M., [127](#)
Cronbach, L. J., [2](#)

Dahler-Larsen, P., [78](#), [292](#)
Davies, R., [61](#)
Dawkins, N., [62e](#)
Deke, J., [204](#)
Diaz, J. J., [270](#)
Dillman, D. A., [43](#), [280](#), [282](#)
Donaldson, S. I., [65](#)
Dornbusch, S. M., [2](#)
Dragoset, L., [204](#)
Dupas, P., [3](#)

Economics Division, Queensland Treasury and Trade, [313](#)
European Commission, Directorate-General for Regional Policy, [313](#)
Evergreen, S. D. H., [297](#)

Fawcett, E. B., [218](#)
Fetterman, D. M., [14](#), [303](#)
Fishbein, M., [67](#)
Fisher, G. A., [104](#)
Fleischer, D. N., [311](#)
Fowler, F. J., [43](#), [273](#)
Fuchs, D., [192](#)
Fuchs, L. S., [192](#)

Fuller, S. C., [269](#)

Fuoco, M. J., [42](#)

Gilbert, J. K., [192](#)

Glass, G. V., [228](#)

Glasser, W., [82](#)

Glick, H. A., [201](#)

Graham, J. W., [282](#)

Greenberg, D. H., [239](#)

Greenseid, L. O., [311](#)

Handa, S., [270](#)

Hanisch-Cerda, B., [261](#)

Harpster- Hagen, A., [3](#)

Hartinger-Saunders, R., [42](#)

Hatry, H. P., [41](#), [133](#)

Hawkins, A. J., [218](#)

Hawthorn, L. R., [133](#)

Hedges, L. V., [232](#)

Heid, C., [268](#)

Heinrich, C. J., [270](#), [273](#), [278](#)

Heinsman, D. T., [182](#)

Henry, G., [123](#), [272](#)

Henry, G. T., [3](#), [38](#), [127](#), [128](#), [267](#), [273](#), [278](#), [310–311](#)

Herbert, J. L., [310](#)

Hess, R. D., [2](#)

Higgins, J. P. T., [232](#)

Holland, P. W., [149](#)

Hollands, F., [261](#)

Hollis-Peel, M. E., [4](#)

Holstein, J. A., [34](#)

Hornik, R. C., [2](#)

House, E. R., [48](#)

Howe, K. R., [48](#)

Hughes, J., [3](#)

Huse, I., [133](#)

Jacob, R., [180](#)
Jarvis, W. J., [3](#)
Johnson, K., [311](#)
Joint Committee on Standards for Educational Evaluation., [307](#)
Julnes, G., [278](#)

Kaftarian, S. J., [14](#), [303](#)
Keehley, P., [139](#)
Kettner, P. M., [68](#), [133](#)
Khan, L. K., [62e](#)
Kho, A., [3](#), [123](#)
King, J. A., [300](#), [311](#)
Kish, L., [38](#)
Kitsuse, J. I., [34](#)
Klein, T. J., [206](#)
Knowlton, L. W., [19](#), [78](#)
Kumar, D. D., [17](#), [34](#), [55](#)

LaBoy, A., [42](#)
Lavenberg, J. G., [4](#)
Lawrence, E., [3](#)
Lawrenz, F., [311](#)
Lemons, C. J., [192](#)
Levin, H. M., [239](#), [261](#)
Leviton, L. C., [62e](#)
Leyland, A., [42](#)
Leyland, E., [269](#)
Lindo, J., [181](#)
Lipsey, M. W., [68](#), [182](#), [232](#)
Liu, X. S., [222](#)
Longmire, L., [139](#)
Luterbach, K. J., [127](#)
Lys, D., [128](#)
Lys, D. B., [127](#)

MacBride, S., [139](#)
McDavid, J. C., [133](#)

McEwan, P. J., [239](#)
McKeganey, N., [42](#)
MacKinnon, D. P., [231](#)
Marchman, J., [3](#)
Mark, M. M., [267](#), [278](#), [310–311](#)
Martin, L. L., [68](#), [133](#)
Mathematica, [293](#)
Medlin, S., [139](#)
Meiers, M. W., [17](#), [34](#)
Melbin, A., [271](#)
Mielke, K.W., [108](#)
Miller, G., [34](#)
Mishan, E. J., [239](#)
Moore-Schiltz, L., [3](#)
Morgan, S. L., [170](#)
Moroney, R., [133](#)
Murphy, K. R., [222](#)
Myors, B., [222](#)

National Commission for the Protection of Human Subjects of
Biomedical and Behavioral Research, [198](#)
Nyborg, K., [242](#)

Ogden, T., [313](#)
O’Neal, K. K., [150](#)

Packard Foundation, [293](#)
Packham, A., [181](#)
Pan, Y., [128](#)
Parker, A., [3](#)
Patriarca, L. A., [127](#)
Patton, M. Q., [14](#), [303](#)
Pauly, M. V., [201](#)
Pawson, R., [19](#), [66](#), [67](#), [68](#)
Pell, J. P., [147](#)
Petrosino, A., [4](#)
Peyton, D. J., [84](#)

Pham, L., [123](#)
Phillips, C. C., [19](#), [78](#)
Phillips, D. C., [2](#)
Piontek, M. E., [297](#)
Potvin, L, [61](#)
Preskill, H., [297](#)
Preston, C., [272](#)
Puig, A., [201](#)
Puma, M., [268](#)

Quah, E., [239](#)

Reardon, S., [204](#)
Rog, D., [62e](#)
Ross, H. L., [228](#)
Rossi, P. H., [65](#), [104](#), [131](#), [279](#)
Rossmo, K., [42](#)
Rothstein, H. R., [232](#)
Routledge, R., [42](#)
Roy, M., [269](#)
Ruel, E., [42](#)
Rutman, L., [82](#), [83](#)

Sanchez, T., [42](#)
Saunders, R. P., [98](#), [99e](#)
Scicchitano, M., [84](#)
Scriven, M., [11](#), [284](#)
Shadish, W. R., [165](#), [182](#)
Shand, R., [261](#)
Sherman, L. W., [194](#)
Skidmore, F., [143](#)
Smit, F., [42](#)
Smith, G. C. S., [147](#)
Smith, M. F., [74](#), [82](#), [83](#)
Smyth, J. D., [43](#), [280](#), [282](#)
Somers, M., [180](#)
Spector, M., [34](#)

Stanley, J. C., [165](#), [166e](#)
Steiner, P. M., [182](#)
Stevahn, L., [300](#)
Stuart, E. A., [172](#)
Swinehart, J. W., [108](#)

Thompson, C. L., [127](#), [272](#)
Thurston, W., [61](#)
Titunik, R., [204](#)
Toal, S. A., [311](#)
Todd, P., [204](#)
Toet, J., [42](#)
Torres, R. T., [297](#)
Treasury Board of Canada, [313](#)
Trouteaud, A., [42](#)
Troxel, A. B., [201](#)
Turpin-Petrosino, C., [4](#)

U.S. General Accounting Office, [36](#), [85](#)

van der Heijden, P., [42](#)
VanderWeele, T. J., [231](#)
Vining, A. R., [239](#)
Visser, Y. L., [17](#), [34](#)
Volkov, B., [311](#)
Volpp, K. G., [201](#)

Wadell, G., [204](#)
Wandersman, A., [14](#), [303](#)
Watkins, R., [17](#), [34](#)
Weimer, D. L., [239](#)
Weiner, J., [201](#)
Weiner, S. S., [2](#)
Weiss, C. H., [8](#), [65](#), [71](#), [73e](#), [74](#)
West, S. L., [150](#)
Wholey, H. P., [61](#), [63](#), [65](#), [82](#), [83](#)
Willis, G., [104](#)
Wilson, D. B., [182](#), [232](#)

Winship, C., [170](#)

Wolach, A., [222](#)

Wright, E. R., [42](#)

Zarling, A., [3](#)

Zhu, P., [180](#)

Zimmer, R., [3](#), [123](#)

Subject Index

Exhibits, figures, and tables are indicated by e, f, or t following the page number. Names that are subjects are in this index. Author names are in a separate author index.

ABA Standards for Criminal Justice: Treatment of Prisoners (ABA), [271](#)
Accelerating Opportunity (AO), [247e](#)
Acceptance and commitment theory (ACT), [3](#)
Accessibility, [110–111](#), [112e](#)
Accountability, process monitoring for, [101–102](#), [101e](#)
Accounting perspectives, [239](#), [244–250](#), [247–249e](#)
Accounting prices, [252–253](#)
Ad hoc measures, [128–132](#), [132e](#)
Administrative data systems, [98–99](#), [100e](#)
Administrative objectives, [15](#)
Administrative standards, [95](#)
Advisory groups, [306](#)
Agency records, [41](#)
Aggressive behavior, [232](#), [233e](#)
Alpha level, [221](#)
American Community Survey, [38–41](#), [40f](#)
American Evaluation Association, [300](#), [307](#), [311](#)
AO (Accelerating Opportunity), [247e](#)
Articulated program theory, [71–74](#)
Assessment of program process, [19–23](#), [20–21e](#)
Assessment of program theory and design, [18–19](#), [20–21e](#)
Assignment variables, [188–190](#), [189f](#), [204](#)
Assumptions, [33–34](#), [60](#), [242](#), [282](#)
Attrition, [160–161](#)
Autism and employment, [26e](#)

Belgian development cooperation, [64e](#)

Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research), [198](#)

Benchmarking, [95](#), [139](#)

Beneficiaries, [294](#)

Benefits. *See* Cost-benefit analysis; Efficiency assessment

Beta (Type II) error, [221–222](#)

Biases

comparison group design and, [159–163](#), [160e](#), [167](#), [177–179](#), [178e](#), [182](#), [208–209](#)

confirmation bias, [5–6](#)

counterfactuals and, [193](#)

repeated measures and, [129](#)

service utilization and, [104–106](#), [105–106e](#), [108–109](#)

See also Selection bias

Black-box evaluation, [87](#)

Blood pressure monitoring, [240–242](#), [241e](#)

Book vending machines, [111](#), [112e](#)

Boundaries of programs, [74–75](#)

Briefings, [283–286](#)

British Breathalyzer crackdown, [228](#)

Budget allocation method, [252](#)

Burger, Warren E., [197e](#)

Campbell, Donald, [166e](#)

Cancer survivors, [52–53](#), [53e](#)

Capture-recapture methods, [42e](#)

Capture tokens, [42e](#)

Case studies, [272–273](#)

Causal design, [272](#). *See also* Comparison group design; Randomized control design; Regression discontinuity design

Causal relationships, [142](#)

Censuses, [41–43](#), [42–43e](#). *See also* U.S. Census

Chance factors, [216](#), [220](#)

Charter schools, [105e](#)

Claremont Graduate University, [302](#)

Client satisfaction, [134](#), [135e](#)

“Cloning,” [172](#), [173e](#)

- Cluster randomized trials, [195](#), [224](#), [225e](#)
- Cluster samples, [39e](#)
- Cohort design, [176–177](#), [178e](#)
- Collaborative (participatory) evaluation, [13](#), [294](#)
- Collaborators in assessment, [80](#), [82–83](#)
- Common good and equity principle, [309e](#)
- Communication, [283–286](#), [295–297](#)
- Community accounting perspective, [246–250](#), [247e](#), [249e](#)
- Comparative interrupted time series design, [177](#), [180](#)
- Comparative needs assessment, [34](#)
- Comparison group design
 - overview, [157–159](#)
 - advantages of, [164–165](#)
 - bias in, [159–163](#), [160e](#), [167](#), [177–179](#), [178e](#), [182](#), [208–209](#)
 - cautions about using, [182–183](#)
 - cohort design, [176–177](#), [178e](#)
 - comparative interrupted time series design, [177](#), [180](#)
 - control group design compared, [158](#)
 - covariate-adjusted, regression-based estimates, [167–171](#), [169e](#)
 - difference-in-differences design, [177–179](#)
 - fixed effects design, [180–181](#)
 - interrupted time series design, [176–181](#), [178e](#)
 - matching and, [171–176](#), [173–175e](#)
 - naive program effect estimates, [165–167](#)
 - planning and, [272](#)
 - propensity scoring, [172–176](#), [174–175e](#)
 - when appropriate, [208](#)
- Comparison groups, [158](#), [186](#)
- Competence principle, [308e](#)
- Conceptualization. *See* Program theory
- Confirmation bias, [5–6](#)
- Consultants, [306](#)
- Contraception, [33](#)
- Control group design
 - overview, [185–186](#)
 - comparison group design compared, [158](#)
 - equivalence and, [187](#)

- ethics of, [196–198](#), [197e](#)
- key concepts in, [190–196](#), [192e](#)
- practical considerations for, [198–199](#), [200–203e](#)
- randomized control design, [187–188](#), [196–199](#), [197e](#), [200–203e](#), [205–209](#), [272](#)
- regression discontinuity design, [188–190](#), [189f](#), [194](#), [199](#), [203–205](#), [206–207e](#), [207–209](#), [272](#)
- types of, [186–190](#), [189f](#)
- Control groups, [186–187](#)
- Convenience samples, [273–274](#)
- Cook, Thomas D., [166e](#)
- Corruptibility of indicators, [134](#)
- Cost analysis, [25](#)
- Cost-benefit analysis
 - accounting perspectives, [244–250](#)
 - comparison of costs and benefits, [256–258](#), [257e](#)
 - cost data collection, [244](#)
 - cost-effectiveness analysis compared, [238](#), [242–243](#)
 - cost estimation, [252–253](#)
 - data needed to calculate, [109](#)
 - defined, [25](#)
 - discounting and, [255](#), [256e](#)
 - distributional effects and, [246](#), [254](#)
 - ex post analyses, [258](#), [259e](#)
 - HIV/AIDS reduction programs, [240e](#)
 - measurement of, [250–253](#)
 - monetizing value of benefits, [242–243](#), [251–252](#)
 - secondary effects and, [246](#), [253–254](#)
 - vacuum effects and, [254](#)
- Cost-effectiveness analysis
 - conducting of, [260](#), [261–262e](#)
 - cost-benefit analysis compared, [238](#), [242–243](#)
 - defined, [25](#)
 - ex post analyses, [258](#)
 - weight control intervention, [243e](#)
- Costs, [110](#), [238](#). *See also* Cost-benefit analysis; Efficiency assessment
- Counterfactuals

- defined, [23](#), [147](#)
- impact evaluation and, [147–149](#)
- missing data problem and, [152–154](#), [274](#), [282](#)
- types of, [191–193](#), [192e](#)
- See also* Comparison group design; Control group design
- Covariate-adjusted, regression-based estimates of program effects, [167–171](#), [169e](#)
- Covariates, [168](#), [170–171](#), [176](#), [222](#)
- Coverage, [104–108](#), [105–106e](#)
- Creaming, [104](#), [105e](#)
- Crime victimization, [50](#), [51e](#)
- Cronbach's alpha, [129](#)
- Crossovers, [194](#)
- Culturally responsive evaluation, [294](#)
- Current Population Survey, [38](#)
- Customer satisfaction, [134](#), [135e](#)
- Cutting-point variables, [188–190](#), [189f](#), [194](#), [204](#)

- D.A.R.E. program, [150](#)
- Data, outcome data and selection bias, [161](#)
- Data analysis plans, [281–283](#)
- Database construction, [281](#)
- Data collection, [41–45](#), [42–44e](#), [103](#), [107](#), [279–281](#)
- Data sharing agreements, [281](#)
- Data sources
 - cost-benefit analysis, [244](#)
 - existing, [38–41](#), [40e](#)
 - outcomes and, [127–128](#)
 - primary data, [277–281](#)
 - secondary data, [281](#)
- Delivery systems, [109–111](#)
- Demand versus need, [49](#)
- Demonstration programs, [10](#)
- Department of Education, [204](#), [266](#), [267e](#), [298](#), [313](#)
- Department of Labor, [108](#)
- Descriptive design, [272](#)
- Diabetes, [100e](#)

Difference-in-differences design, [177–179](#)
Differential attrition, [160–161](#)
Differential effects, [143](#)
Direct instrumental use of results, [310–312](#)
Direct observations, [86–87](#)
Direct outcomes, [68](#)
Direct target populations, [47](#)
Disaggregation of data, [137](#)
Discounting, [255](#), [256e](#)
Discount rates, [255](#)
Displacement, [253](#)
Dissemination of results, [296–297](#)
Distal outcomes, [68](#), [121](#), [122e](#)
Distributional effects, [246](#), [254](#)
Diversity, [291](#), [301–306](#)
Dose delivered, [99e](#)
Dose received, [99e](#)
Dose-response analysis, [143](#), [228](#)
Drop-out rates, [108–109](#), [161](#)
Drug urine screens, [257–258](#), [257e](#)

Econometric estimation method, [251](#)
Economic Report of the President, [313](#)
Education of evaluators, [302–303](#)
Effectiveness evaluation, [190](#)
Effective sample size, [224](#)
Effect size statistic, [213](#), [214–215e](#)
Efficacy evaluation, [190](#)
Efficacy trials, [145](#), [190](#)
Efficiency assessment
 overview, [237–239](#)
 accounting perspectives and, [239](#)
 comparative utility perspective on, [239](#), [240e](#), [242](#)
 defined, [25](#)
 ex ante analyses, [239–242](#), [241e](#)
 ex post analyses, [239](#), [242](#), [258](#), [259e](#)
 when appropriate, [27](#)

See also Cost-benefit analysis; Cost-effectiveness analysis

Elite evaluation organizations, [306–307](#)

Emergency room urine drug screens, [257–258](#), [257e](#)

Employment programs, [106](#), [106e](#), [247e](#), [249e](#)

Empowerment evaluation, [13](#), [294](#)

Enlightenment, [73e](#), [310](#)

Equal treatment under the law, [198](#)

Equity principle, [309e](#)

Ethical Guiding Principles (AEA), [307–310](#), [308–309e](#)

Ethics

- counterfactuals and, [193](#)
- data collection, [279](#)
- Ethical Guiding Principles*, [307–310](#), [308–309e](#)
- multiple stakeholders and, [295](#)
- of random assignment, [196–198](#), [197e](#)

Ethics codes, [310](#)

Evaluability assessment, [61–63](#), [62–64e](#)

Evaluation. *See* Program evaluation

Evaluation field, [234](#)

Evaluation influence, [310–311](#)

Evaluation organizations, [300](#), [301e](#)

Evaluation partnerships, [298](#)

Evaluation plans

- overview, [265–266](#), [267e](#)
- communication about, [283–286](#)
- data analysis and, [281–283](#)
- data collection and management and, [276–281](#)
- evaluation questions and, [269–271](#), [270e](#)
- measures or observations used, [274–276](#)
- project management plans and, [286–287](#), [287–289e](#)
- purpose and scope of, [266–276](#), [270e](#)
- qualitative versus quantitative methods and, [304–305](#)
- research design and, [272–273](#)
- samples and, [273–274](#)

Evaluation questions

- assessment of program process, [19–23](#), [22e](#)
- assessment of program theory and design, [18–19](#), [20–21e](#)

- cost analysis and efficiency assessment, [25–27](#), [26e](#)
- data analysis and, [282](#)
- evaluability assessment and, [63](#)
- impact evaluation and, [23–25](#), [24–25e](#)
- interplay among, [27–29](#)
- needs assessment and, [17](#), [18e](#)
- planning and, [269–271](#), [270e](#)
- role of, [10–16](#)
- types of, [16–17](#)

Evaluation reports, [283–285](#), [296–297](#)

Evaluation results, use of, [310–312](#)

The Evaluation Society (Dahler-Larsen), [292](#)

Evaluation sponsors, [9](#), [13](#). *See also* Stakeholders

Evaluation time versus political time, [298](#)

Evaluators

- definition lacking for, [300](#)
- diversity among, [301–306](#)
- education of, [302–303](#)
- role of, [297](#), [305–306](#)
- specialization by, [302](#)

Evaluator's Ethical Guiding Principles (AEA), [307–310](#), [308–309e](#)

Evaluator-stakeholder relationship, [12–14](#)

Evidence-based practice movement, [312–313](#)

Evolutionary epistemology, [166e](#)

Exact matching, [172](#), [173e](#)

Ex ante efficiency analyses, [239–242](#), [241e](#)

Experimental and Quasi-Experimental Design for Research (Campbell & Stanley), [165](#), [166e](#)

Expert review panels, [82–83](#)

Experts, [306](#)

Ex post efficiency analyses, [239](#), [242](#), [258](#), [259e](#)

Externalities (secondary effects), [246](#), [253–254](#)

External validity, [154](#), [164](#), [166e](#), [181](#)

Face validity, [129](#)

Fairness, [203](#). *See also* Ethics

Federal Judicial Center, [197–198](#), [197e](#)

Feedback, process monitoring as, [98–99](#), [100e](#)
Feeling Good television program, [108](#)
Fidelity, [92](#), [98](#), [99e](#), [144](#), [146](#)
Final reports, [284–285](#), [296–297](#)
Fiscal records, [244](#), [245e](#)
Fixed effects design, [180–181](#)
Flourishing Children Project, [132e](#)
Focus groups, [53–55](#), [54e](#)
Forcing variables, [188–190](#), [189f](#)
Forecasting of needs, [45](#)
Form and language for reports, [296–297](#)
Formative evaluation, [11](#)
Full logic models, [71](#), [73e](#)
Fundamental problem of causal inference, [152–154](#)
Funding, [101–102](#), [101e](#), [120](#)
Funding allocation method, [252](#)

GAO (U.S. Government Accountability Office), [2](#), [102](#)
General Accounting Office, [36](#), [85](#)
Goals, [36](#), [77](#), [120](#)
Government Accountability Office (GAO), [2](#), [102](#)
GREAT program, [82e](#)
Guidance Document on Monitoring and Evaluation (Directorate-General for Regional Policy), [313](#)
Guiding Principles for Evaluators (AEA), [307](#)

Hawthorne effect, [199](#)
Head Start evaluation, [268](#), [269](#)
Health care provider training in India, [24–25e](#)
Health insurance in Peru, [206–207e](#)
Hearing protection, [229–231](#), [230e](#)
HEAT, [138e](#)
Higher education access tracking, [138e](#)
High Fidelity Wraparound (Wrap), [248e](#)
High school completion rates, [260](#), [261–262e](#)
HIV/AIDS, [18e](#), [135e](#)
Holdback samples, [279–280](#)

Homelessness, [7e](#), [33](#), [41–42](#), [42e](#)
Household Targeting Index, [206–207e](#)
Hypothetical question method, [251–252](#)

ICC (intraclass correlation coefficients), [224](#), [225e](#)

Immediate outcomes, [68](#)

Impact evaluation

- overview, [141–142](#), [185–186](#)
- choice of design for, [205–209](#)
- as comparative, [186](#)
- counterfactuals and, [23](#), [147–149](#)
- defined, [23](#)
- evaluation questions and, [23–25](#), [24–25e](#)
- fundamental problem of causal inference, [152–154](#)
- importance of, [142](#)
- key concepts in, [190–196](#), [192e](#)
- meta-analysis and, [231–234](#), [233e](#)
- policy space and, [299–300](#)
- potential outcomes framework and, [149–152](#), [150–151t](#)
- process evaluation and, [23](#), [92](#), [98](#), [101](#)
- program effects and, [119](#), [212](#)
- questions to ask, [143–144](#), [145e](#)
- validity and, [153–154](#)
- when appropriate, [25](#), [144–147](#)

See also Comparison group design; Control group design

Impacts, [119](#), [158](#). *See also* Program effects

Impact theory

- overview, [65–68](#), [66e](#), [69e](#)
- needs assessment and, [81](#)
- outcomes and, [121](#), [122e](#)
- problems with, [87](#)

Implementation. *See* Process evaluation

Implementation failure, [28](#)

Implementation fidelity, [92](#), [98](#), [99e](#), [144](#), [146](#)

Implicit program theory, [74](#)

Incentives, [199](#), [200–201e](#)

Incidence, [50](#)

Incidence rates, [50](#), [51e](#)
Income Maintenance Experiments, [143](#)
Independent evaluation, [13–14](#)
Independent institutional review boards, [279](#)
Indirect target populations, [47](#)
Individual-target population accounting perspective, [246](#), [247e](#), [249e](#)
Infectious disease, [3](#)
Inflation adjustment, [255](#)
Influence, [266–267](#)
Information versus judgments, [294](#)
Informed consent, [310](#)
Insensitivity of measures, [130–131](#)
Insider status, [66–67](#), [305](#)
Institute of Education Sciences, [266](#), [267e](#), [298](#)
Institutional review boards, [279](#)
Intake data, [136–137](#)
Integrity principle, [308–309e](#)
Intent-to-treat (ITT) effects, [194](#)
Interaction terms, [226](#)
Interfering events, [162](#)
Interim reports, [284–285](#)
Internal consistency reliability, [129](#)
Internal rates of return, [255](#)
Internal validity, [154](#), [164](#), [166e](#), [179](#)
International Organization for Cooperation in Evaluation (IOCE), [301e](#)
Interrupted time series design, [176–181](#), [178e](#)
Intervention groups, [158](#)
Intraclass correlation coefficients (ICC), [224](#), [225e](#)
IOCE (International Organization for Cooperation in Evaluation), [301e](#)
ITT (intent-to-treat) effects, [194](#)
iZone schools, [3](#)

Janus variables, [78–79](#)
Joint Committee on Standards for Educational Evaluation, [307](#)
Judgments versus information, [294](#)
Juvenile delinquency, [124](#), [125e](#), [259e](#)

Key informants, [44e](#), [45f](#), [55](#)
Kindergarten Peer-Assisted Learning Strategies (K-PALS), [191–193](#), [192e](#)
Knowledge generation, [12](#)

Language and form for reports, [296–297](#)
Law of unintended consequences, [142](#)
Leadership academies, [96–98](#), [97e](#)
Legal standards, [95](#)
LGBTQ community, [33](#), [42e](#)
Literature reviews
 meta-analysis and, [231–234](#), [233e](#)
 outcomes and, [122–123](#)
 program theory and, [84–85](#)
 sensitivity of measures, [131](#)
Logic, [82–84](#)

Magnitude of program effects, [212–213](#), [214–215e](#), [215](#)
Mail surveys, [42–43](#), [43e](#)
Malaria, [2–3](#)
Management-oriented process evaluation, [102–103](#)
Market prices, [252](#)
Market valuation method, [251](#)
Masking, [227](#)
Massachusetts health insurance reforms, [177–179](#), [178e](#)
Matching, [163](#), [171–176](#), [173–175e](#)
Maturation, [163](#)
MDES (minimum detectable effect size), [216](#), [223e](#)
Measures used, [274–276](#)
Mediator analysis, [229–231](#)
Mediator variables, [229](#)
Memorandum of understanding for data sharing, [281](#)
Mendota Juvenile Treatment Center (MJTC), [259e](#)
Meta-analysis, [218](#), [231–234](#), [233e](#)
Milestones, [287](#), [287–289e](#)
Milwaukee County public health priorities, [44e](#)
Minimum detectable effect size (MDES), [216](#), [223e](#)

Missing data problem, [152–154](#), [274](#), [282](#)
Mixed data, [277–279](#)
Mixed-method design, [278–279](#), [304–305](#)
Mixed-mode surveys, [280–281](#)
Moderator analysis, [226–229](#)
Moderator variables, [226](#)
Money measurement method, [251](#)
Monitoring and evaluation (M&E), [92](#). *See also* Process evaluation;
Process monitoring
Mosquito netting, [2–3](#)
Motivation, for evaluation, [9–10](#)
Multiple interventions, [195–196](#)
Multiple measures, [124–127](#), [126e](#)
Multistage samples, [39e](#)
Multivariate regression, [168](#), [170–171](#)

Naive program effect estimates, [165–167](#)
National Survey of Household Drug Use, [41](#)
Needs assessment
 comparative needs assessment, [34](#)
 coverage and, [106](#)
 data collection for, [41–45](#), [42–44e](#)
 defined, [17](#), [32](#)
 evaluation questions and, [17](#), [18e](#)
 evaluator role, [32–34](#)
 existing data sources, [38–41](#), [40e](#)
 extent of problem, determination of, [37–45](#), [39–40e](#), [42–44e](#)
 geographic boundaries of, [46](#)
 importance of, [31–32](#)
 needs description, [52–55](#), [53–54e](#)
 phases of, [35e](#), [37](#)
 problem definition, [34–37](#), [36e](#)
 program theory and, [80–81](#)
 target population identification, [46–48](#)
Need versus demand, [49](#)
Negative side effects, [143](#)
Net benefits, [255](#)

Net effects, [23](#)
Net rate of return, [255](#)
Nice-to-know measures, [276](#)
Noise protection devices, [229–231](#), [230e](#)
Noncompleters, [108–109](#), [161](#)
Noncompliance, [194](#)
Nonrandomized design. *See* Comparison group design
North Carolina's Race to the Top, [269](#), [270e](#)

Observational studies. *See* Comparison group design

Observations, [86–87](#)
Odds ratio, [213](#), [214–215e](#)
Office of Administration and Budget, [102](#)
Offshoring and tax reductions, [33](#)
Opportunity costs, [253](#)

Organizational plans
 overview, [68–71](#), [71e](#)
 defined, [67](#)
 process evaluation and, [93](#)
 process theory and, [109–111](#), [112e](#)

Organizations for evaluation, [300](#), [301e](#)

Outcome changes, [117–119](#), [118e](#)

Outcome data, selection bias and, [161](#)

Outcome level, [117–119](#), [118e](#)

Outcome measurement, [124–132](#), [125–126e](#), [132e](#), [217](#), [276](#)

Outcome monitoring
 defined, [21](#), [93](#), [133](#)
 indicators for, [133–134](#), [135e](#)
 interpretation of data, [136–139](#), [138e](#)
 multiple influences on, [117](#), [133](#)
 pitfalls in, [134–136](#)

Outcomes
 overview, [115–116](#)
 client satisfaction, [134](#), [135e](#)
 defined, [116–117](#)
 identification of candidates for measurement, [119–123](#), [120e](#), [122e](#)

- impact theory and, [121](#), [122e](#)
- interpretation of, [136–139](#), [138e](#)
- measurement of, [124–132](#), [125–126e](#), [132e](#)
- practical significance and, [15–16](#)
- proximal versus distal, [68](#), [121](#), [122e](#)
- terminology, [117–119](#), [118e](#)

Outputs, [116](#)

Outsider status, [66–67](#), [305](#)

Overcoverage, [104–105](#), [106](#)

Pareto criterion, [254](#)

Participant study samples, [154](#)

Participatory (collaborative) evaluation, [13](#), [294](#)

Performance criterion, [15–16](#), [16e](#)

Permission for data collection, [279](#)

Personnel, [286](#)

Perverse incentives, [134](#)

Pilot programs, [145](#), [190](#)

Planning stage. *See* Evaluation plans

Plausibility, [82–84](#)

Policy on Results (Treasury Board of Canada), [313](#)

Policy significance, [299–300](#)

Policy space, [299–300](#)

Political time versus evaluation time, [298](#)

Politics

- administrative objectives and, [15](#)
- context for evaluation, [9–10](#), [297–300](#)
- counterfactuals and, [191–193](#), [192e](#)
- hidden agendas, [12](#)

Population at risk, [49](#)

Population in need, [49](#)

Potential outcomes framework, [149–152](#), [150–151t](#)

Potential Pareto improvement, [254](#)

Poverty, [34–36](#), [36e](#), [38–41](#), [40e](#)

Predictive validity, [130](#)

Pre-post comparisons, [137](#)

Present value, [255](#), [256e](#)

Pretests, [222](#)
Prevalence, [50](#)
Prevalence rates, [50](#)
Primary data, [277](#), [279–281](#)
Primary dissemination, [283–285](#), [296–297](#)
Privacy, [55](#)
Probability, [142](#)
Probability sampling, [37–38](#), [39e](#), [273](#), [279–280](#)
Problem definition, [34–37](#), [36e](#)
Process evaluation
 overview, [91–93](#)
 components of, [99e](#)
 criteria for, [94–96](#)
 defined, [92](#)
 forms of, [96–99](#), [97e](#), [99–100e](#)
 impact evaluation and, [23](#), [92](#), [98](#), [101](#)
 management-oriented process evaluation, [102–103](#)
 organizational plans and, [109–111](#), [112e](#)
 process monitoring and, [100–103](#), [101e](#)
 service utilization and, [103–109](#), [105–106e](#)
Process monitoring, [92–93](#), [98–103](#), [100–101e](#)
Process theory
 defined, [67](#)
 function assessment and, [109](#)
 needs assessment and, [81](#)
 problems with, [87](#)
 process evaluation and, [94](#)
Product-moment correlation, [129](#)
Professionalism, [291–292](#), [307–310](#), [308–309e](#)
Professional schools, [302–303](#)
Program effects
 overview, [185–186](#), [211–212](#)
 bias in estimation of, [159–163](#), [160e](#)
 criteria for, [14–16](#), [16e](#)
 defined, [117–119](#), [118e](#), [141](#), [158](#)
 impact evaluation and, [119](#), [193–194](#)
 magnitude of, [212–213](#), [214–215e](#), [215](#)

mediator analysis and, [229–231](#), [230e](#)
moderator analysis and, [226–229](#)
potential outcomes framework and, [152](#)
practical significance and, [15–16](#), [216–220](#), [220–221e](#), [300](#)
relativity of, [5](#)
See also Comparison group design; Control group design; Impact headings

Program effect size

estimates of, [215–216](#)
magnitude of, [212–213](#), [214–215e](#), [215](#)
minimum detectable effect size, [216](#), [223e](#)
practical significance and, [216–220](#), [220–221e](#)
statistical power and, [221–226](#), [223e](#), [225e](#)
statistical significance and, [220–221](#)
valence and, [215](#)

Program evaluation

defined, [1–3](#), [6](#)
effectiveness of programs, [9](#)
elements of, [8e](#)
future of, [312–313](#)
need for, [3–4](#)
political and organizational context, [9–10](#)
purposes of, [9–12](#)
social ecology of, [292–300](#)
social research methods and, [6–8](#), [8e](#)
as systematic, [4–6](#)

Program Evaluation and Methodology Division, [85](#)

The Program Evaluation Standards: A Guide for Evaluators and Evaluation Users (Joint Committee), [307](#)

Program groups, [158](#)

Program impact evaluation. *See* Impact evaluation

Program impacts. *See* Impacts

Program impact theory. *See* Impact theory

Program monitoring, [21](#)

Program outcomes. *See* Outcomes

Program outputs, [116](#), [276](#)

Program process

- defined, [67](#)
- evaluation questions and, [19–23](#), [22e](#)
- impact evaluation and, [28](#), [146](#)
- outcomes and, [136–137](#), [138e](#)
- See also* Process evaluation
- Program process theory. *See* Process theory
- Program records, [107](#)
- Program sponsor accounting perspective, [246](#), [248e](#), [249e](#)
- Program theory
 - overview, [59–61](#)
 - articulated program theory, [71–74](#)
 - assessment of, [79–87](#)
 - defined, [60](#)
 - description of, [65–66e](#), [65–71](#), [69–71e](#), [78–79](#)
 - eliciting of, [71–79](#), [72–73e](#), [75–76e](#)
 - evaluability assessment and, [61–63](#), [62–64e](#)
 - explicating of, [75–76e](#), [75–77](#)
 - full logic models, [71](#), [73e](#)
 - impact evaluation and, [146](#)
 - impact theory and, [65–68](#), [66e](#), [69e](#)
 - organizational plans and, [67](#), [68–71](#), [71e](#), [96](#), [109–111](#), [112e](#)
 - outcomes from assessment of, [87](#)
 - process evaluation and, [94](#)
 - reality versus, [78](#)
 - service utilization plans, [67](#), [68](#), [70e](#), [103–109](#), [105–106e](#)
- Project management plans, [286–287](#), [287–289e](#)
- Proof of concept, [145](#), [190](#)
- Propensity scoring, [172–176](#), [174–175e](#)
- Proposals. *See* Evaluation plans
- Protocols, [277](#)
- Proximal outcomes, [68](#), [121](#), [122e](#), [229–231](#), [230e](#)
- Purpose and scope of evaluation, [266–276](#), [270e](#)
- Purposes of evaluation, [9–12](#)

- Qualitative data, [277–279](#), [282](#)
- Qualitative methods, [52–55](#), [53–54e](#), [304–305](#)
- Quality-adjusted life-years (QALYs), [243e](#)

Quantitative assignment variables, [188–190](#), [189f](#)
Quantitative data, [277–279](#), [282](#)
Quantitative methods, [304–305](#)
Quasi-experiments, [165](#). *See also* Comparison group design
Queensland Government Program Evaluation Guidelines (Economics Division), [313](#)

Race to the Top, [269](#), [270e](#)
Random assignment, [187–188](#), [196–199](#), [197e](#), [200–203e](#), [273](#)
Randomized control design
 overview, [187–188](#)
 ethics and, [196–198](#), [197e](#)
 planning and, [272](#)
 practical considerations for, [198–199](#), [200–203e](#)
 when to use, [205–209](#)
Random sampling, [37–38](#), [39e](#)
Rate of return, [255](#)
Rates of incidence and prevalence, [50](#), [51e](#)
Reach, [99e](#)
Reading time, [202–203e](#)
Recommendations, [284–285](#)
Record keeping, [107](#)
Recruitment, [99e](#)
Redesign, after program theory assessment, [87](#)
Regional leadership academies (RLA), [96–98](#), [97e](#)
Registries, [313](#)
Regression, [174](#)
Regression discontinuity design
 overview, [188–190](#), [189f](#)
 application of, [199](#), [203–205](#), [206–207e](#)
 intent-to-treat effects and, [194](#)
 planning and, [272](#)
 when appropriate, [207–209](#)
Regression to the mean, [163](#), [179](#)
Relative Risk program, [240e](#)
Relativity of program effects, [5](#)
Release of findings, [283–286](#), [296–297](#)

Reliability of data, [278](#)
Reliability of measures, [128–129](#), [132e](#)
Repeated measures, [129](#)
Reports, [283–285](#), [296–297](#)
Research
 meta-analysis and, [231–234](#), [233e](#)
 needs assessment and, [82e](#), [84–85](#)
 outcomes and, [122–123](#)
 sensitivity of measures, [131](#)
Research-practitioner partnerships, [298](#)
Resource allocation issues. *See* Cost-benefit analysis; Efficiency assessment
Resource planning, [286–287](#)
Respect principle, [309e](#)
Rossi, Peter H., [6](#), [7e](#)
Rubin, Donald, [149](#)

Samples, [37–38](#), [39e](#), [273–274](#), [279–280](#)
Sample size, [220](#), [222](#), [223e](#), [224](#), [273](#)
Sampling error, [220–221](#)
Sampling frames, [38](#)
Satisfaction, [99e](#)
Scared Straight program, [4](#)
School underperformance, [75–76e](#), [78–79](#), [95–96](#)
Secondary data, [277](#)
Secondary dissemination, [296–297](#)
Secondary effects (externalities), [246](#), [253–254](#)
Secular trends, [162](#)
Selection bias
 comparison group design and, [159–161](#), [160e](#)
 effects of, [148–149](#)
 interrupted time series design, [181](#)
 naive program effect estimates and, [167](#)
 propensity scoring and, [175](#)
 statistical model to avoid, [205](#)
 subgroup definition and, [227](#)
 See also Control group design

Self-selection, [104](#)
Sensitivity of measures, [130–131](#)
Serious emotional disturbance wraparound services, [248e](#)
Service utilization, outcomes and, [136–137](#), [138e](#)
Service utilization plans
 overview, [68](#), [70e](#)
 assessment of, [103–109](#), [105–106e](#)
 defined, [67](#)
 process evaluation and, [93](#)
Shadish, William, [166e](#)
Shadow prices, [252–253](#)
Simple random samples, [39e](#)
Site visits, [277](#), [280](#)
Smoking cessation, [200–201e](#)
Snowball sampling, [55](#)
Social accounting perspective, [246–250](#), [247e](#), [249e](#)
Social ecology of evaluations, [292–300](#)
Social indicators, [38–41](#), [40e](#)
Social research methods, application of, [6–8](#), [8e](#)
South Central Economic Development District, [43e](#)
Specificity, [80–81](#), [120](#), [120e](#)
Speed cameras, [174–175e](#)
Sponsors, [101–102](#), [101e](#)
Stakeholders
 accounting perspectives and, [250](#)
 defined, [9](#)
 efficiency assessment and, [238](#)
 evaluator orientation towards, [303–304](#)
 evaluator relationship with, [12–14](#)
 findings and, [283](#)
 influence and, [266–267](#)
 multiple, [293–296](#)
 outcome identification and, [120](#)
 outcome monitoring and, [139](#)
 standards and, [271](#)
 target population identification and, [48](#)
 theory description and, [79](#)

- time issues, [298](#)
- types of, [13](#)
- Standardized mean difference, [213](#), [214e](#)
- Standards, [271](#), [307–310](#), [308–309e](#)
- Standards for Criminal Justice: Treatment of Prisoners* (ABA), [271](#)
- Stanley, Julian, [166e](#)
- Statistical criteria, [15–16](#)
- Statistical noise, [216](#), [220](#)
- Statistical power, [221–226](#), [223e](#), [225e](#)
- Statistical significance, [220–221](#)
- Statistical significance testing, [220–221](#)
- Stratification, [174](#)
- Stratified random samples, [39e](#)
- Subgroup effects, [226–229](#)
- Successive approximation, [75](#)
- Summative evaluation, [11–12](#)
- Support functions, [111](#)
- Survey of Income and Program Participation, [38](#), [108](#)
- Surveys, [41–43](#), [42–43e](#), [107–108](#), [280–281](#)
- Switchers, [181](#)
- Systematic evaluation. *See* Program evaluation
- Systematic information, importance of, [32–33](#)
- Systematic inquiry principle, [308e](#)
- Systematic samples, [39e](#)

- Tacit theory, [74](#)
- Talent Search, [261–262e](#)
- Targeted programs, [47](#)
- Target populations
 - defined, [32](#)
 - description of, [49–50](#), [51e](#)
 - identification of, [46–48](#), [104](#)
- Tax reductions and offshoring, [33](#)
- Technical review committees, [306](#)
- Teen pregnancy, [101e](#)
- Test-retest reliability, [129](#)
- Theories of change, [73e](#), [75–76e](#). *See also* Impact theory

Theory. *See* Program theory
Theory assessment, [79–87](#)
Theory failure, [28](#)
Theory of action, [75–76e](#)
Time issues, [283](#), [287](#), [287–289e](#), [298](#)
Transitional Housing Services for Victims of Domestic Violence: A Report From the Housing Committee of the National Task Force to End Sexual and Domestic Violence (Correia & Melbin), [271](#)
Treatment-on-the-treated (TOT) effects, [194](#)
Two-step program impact theory, [68](#)
Type I error, [222](#)
Type II (beta) error, [221–222](#)

Ultimate outcomes, [68](#)
Unanticipated effects, [123](#), [143](#)
Unavoidable missing data problem, [152–154](#), [274](#), [282](#)
Undercoverage, [104](#), [106](#)
Uniform Subsidy program, [240e](#)
Unintended effects, [123](#), [142](#)
Universal programs, [47](#)
Urine drug screens, [257–258](#), [257e](#)
U.S. Census, [38–41](#), [40f](#)
U.S. Department of Education, [204](#), [266](#), [267e](#), [298](#), [313](#)
U.S. Department of Labor, [108](#)
U.S. General Accounting Office, [36](#), [85](#)
U.S. Government Accountability Office (GAO), [2](#), [102](#)
Utilization-focused evaluation, [13](#)

Vacuum effects, [254](#)
Valence, [215](#), [274](#)
Validation studies, [182](#)
Validity, [129–130](#), [132e](#), [153–154](#)
Variables
 assignment variables, [188–190](#), [189f](#), [204](#)
 cutting-point variables, [188–190](#), [189f](#), [194](#), [204](#)
 forcing variables, [188–190](#), [189f](#)
 Janus variables, [78–79](#)

mediator variables, [229](#)
moderator variables, [226](#)
quantitative assignment variables, [188–190](#), [189f](#)
Vocabulary, [297](#)

Weight control intervention, [243e](#)
Weighting, [174](#), [254](#)
Weiss, Carol, [71](#), [72e](#), [310](#)
West End Walkers 65+, [20–21e](#)
Western Michigan University, [302](#)
What Works Clearinghouse, [204](#), [313](#)
Within-study comparisons, [182](#)
Wrap (High Fidelity Wraparound), [248e](#)

Youth Empowerment Solutions (YES), [22e](#)