Kenneth J. Linfield
and Emil J. Posavac



PROGRAM EVALUATION

METHODS AND CASE STUDIES

EDITION 9

eRESOURCES

ROUTLEDGE

Program Evaluation

This text provides a solid foundation in program evaluation, covering the main components of evaluating agencies and their programs, how best to address those components, and the procedures to follow when conducting evaluations. Different models and approaches are paired with practical techniques, such as how to plan an interview to collect qualitative data and how to use statistical analyses to report results. In every chapter, case studies provide real world examples of evaluations broken down into the main elements of program evaluation: the needs that led to the program, the implementation of program plans, the people connected to the program, unexpected side effects, the role of evaluators in improving programs, the results, and the factors behind the results. In addition, the story of one of the evaluators involved in each case study is presented to show the human side of evaluation.

The Ninth Edition offers enhanced and expanded case studies, making them a central organizing theme. The new edition also features online resources with an instructor test bank, sample program evaluation reports, and additional annotated resources.

Kenneth J. Linfield is Professor and Director of Graduate Training in the School of Professional Psychology at Spalding University in Louisville, Kentucky, USA where he teaches the Program Evaluation course in the doctoral Clinical Psychology program.

Emil J. Posavac is Professor Emeritus of Psychology at Loyola University of Chicago, USA where he served as director of the Applied Social Psychology Graduate Program and chairman of the Psychology Department.

Program Evaluation
Methods and Case Studies
Ninth Edition

Kenneth J. Linfield and Emil J. Posavac

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

Ninth edition published 2019
by Routledge
711 Third Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

© 2019 Taylor & Francis

The right of Kenneth J. Linfield and Emil J. Posavac to be identified as authors of this work has been asserted by them in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this book may be reprinted or reproduced or utilised in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

First edition published by Prentice Hall, Inc. (1980)
Eighth edition published by Pearson Education, Inc. (2011)

Library of Congress Cataloging-in-Publication Data

Names: Linfield, Kenneth J., author. | Posavac, Emil J., author.
Title: Program evaluation : methods and case studies / Kenneth J. Linfield and Emil J. Posavac.
Description: 9th edition. | New York : Routledge, 2018. | Revised edition of Program evaluation, c2011. | Includes bibliographical references and index.
Identifiers: LCCN 2018010276| ISBN 9781138103962 (hardback : alk. paper) | ISBN 9781315102429 (ebook)
Subjects: LCSH: Evaluation research (Social action programs)—United States.
Classification: LCC H62.5.U5 P62 2018 | DDC 361.6/1072—dc23LC record available at https://lccn.loc.gov/2018010276

ISBN: 978-1-138-10396-2 (hbk)
ISBN: 978-1-315-10242-9 (ebk)

Typeset in Times New Roman MT Std
by diacriTech, Chennai

Visit the eResources: www.routledge.com/9781138103962

Ken Linfield dedicates this book to Paul and Lois Linfield, his parents, in memoriam, and to the love of his life, Amy Linfield, who has always believed in him.

Emil Posavac dedicates this book to his grandchildren: Austin, Brett, and Caleigh.

Brief Contents

Preface

1 Program Evaluation: An Overview

2 Planning an Evaluation

3 Developing and Using a Theory of the Program

4 Developing Measures of Implementation and Outcomes

5 Ethics in Program Evaluation

6 The Assessment of Need

7 Monitoring the Implementation and the Operation of Programs

8 Qualitative Evaluation Methods

9 Outcome Evaluations with One Group

10 Quasi-Experimental Approaches to Outcome Evaluation

11 Using Experiments to Evaluate Programs

12 Analyses of Costs and Outcomes

13 Evaluation Reports: Interpreting and Communicating Findings

14 How to Encourage Utilization

References
Author Index
Subject Index

Contents

Preface

1 Program Evaluation: An Overview

Evaluation Areas That Need to Be Considered
- *Meeting Needs*
- *Implementation*
- *Stakeholders*
- *Side Effects*
- *Improvement Focus*
- *Outcomes*
- *Nuances (Mechanisms)*
- *MISSION*

Common Types of Program Evaluations
- *Assess Needs of the Program Participants*
- *Examine the Process of Meeting the Needs*
- *Measure the Outcomes and Impacts of a Program*
- *Integrate the Needs, Costs, and Outcomes*

Important Issues in Evaluations
- *Time Frames of Needs*
- *Extensiveness of Programs*

Purpose of Program Evaluation
The Roles of Evaluators
- *Principles, Competencies, and Credentialing*
- *Comparison of Internal and External Evaluators*
- *Evaluation and Service*
- *Activities Often Confused with Program Evaluation*
- *Evaluation and Related Activities of Organizations*

Summary and Preview
Study Questions
Additional Resource

2 Planning an Evaluation

Overview
Evaluation models
- *An Improvement-Focused Approach*

Steps in Preparing to Conduct an Evaluation
- *Identify the Program and Its Stakeholders*
- *Become Familiar with Information Needs*

Plan the Evaluation
Dysfunctional Attitudes Making an Evaluation Challenging
Our Program Is Exemplary
An Evaluation Will Offend the Staff
An Evaluation Will Inhibit Innovation
Our Program Will Be Terminated
The Information Will Be Misused
Evaluation Methods Are Less Sensitive Than Perceptions of Staff
Evaluation Drains Program Resources
Evaluations Have Little Impact
The Effect of These Attitudes
Summary and Preview
Study Questions
Additional Resource

3 Developing and Using a Theory of the Program

Developing a Program Theory
Why a Program Theory Is Helpful
How to Develop a Program Theory
Implausible Program Theories
Evaluation Questions Flow from Program Theory
Does the Program or Plan Match the Values of the Stakeholders?
Does the Program or Plan Match the Needs of the People to Be Served?
Does the Program as Implemented Fulfill the Plans?
Do the Outcomes Achieved Measure Up to the Goals?
Seeking Information about the Levels of Outcome Expected
Evaluating without a Theoretical Model
Cautions in Choosing Evaluation Criteria
Criteria Must Reflect a Program's Purposes
Criteria Must Be under the Staff's Control
Stakeholders Must Participate in Selecting Evaluation Criteria
Identifying Goals and Objectives
How Much Agreement on Goals Is Needed?
Different Types of Goals
Evaluation Questions That Apply to All Programs
Is the Program Accepted?
Are the Resources Devoted to the Program Being Expended Appropriately?
Some Practical Limitations in Selecting Evaluation Criteria
Evaluation Budget
Time Available for the Project
Criteria That Are Credible to the Stakeholders
Summary and Preview
Study Questions
Additional Resource

4 Developing Measures of Implementation and Outcomes

Sources of Data for Evaluation
Intended Beneficiaries of the Program
Providers of Services

Observers
Which Sources Should Be Used?
Gathering Information Wisely
Use Nonreactive Measures
Use Only Variables Relevant to the Evaluation
Use Reliable Measures
Use Valid Measures
Use Measures That Can Detect Change
Use Cost-Effective Measures
Types of Measures of Evaluation Criteria
Written Surveys and Interviews with Program Participants
Checklists, Tests, and Records
Preparing Special Surveys
Format of a Survey
Instructions and Pretests
Summary and Preview
Study Questions
Additional Resource

5 Ethics in Program Evaluation

Examples of the Problem
A Project That Cannot Be Done Well
Advocacy versus Evaluation
Standards for the Practice of Evaluation
Ethical Issues Involved in the Treatment of People
Detecting Ineffective Programs
Obtaining Informed Consent
Maintaining Confidentiality
Role Conflicts Facing Evaluators
Recognizing the Interests of Different Stakeholders
Program Managers Are Concerned with Efficiency
Staff Members Seek Assistance in Service Delivery
Clients Want Effective and Appropriate Services
Community Members Want Cost-Effective Programs
The Validity of Evaluations
Valid Measurement Instruments
Skilled Data Collectors
Appropriate Research Design
Adequate Descriptions of Program and Procedures
Avoiding Possible Negative Side Effects of Evaluation Procedures
Can Someone Be Hurt by Inaccurate Findings?
Consider Statistical Type II Errors
Pay Attention to Unplanned Negative Effects
Analyze Implicit Values Held by the Evaluator
Institutional Review Boards and Program Evaluation
Ethical Problems Evaluators Report
Summary and Preview
Study Questions
Additional Resource

6 The Assessment of Need

Definitions of Need
Sources of Information for the Assessment of Need
Describing the Current Situation
Social Indicators of Need
Community Surveys of Need
Services Already Available
Key Informants
Focus Groups and Open Forums
Inadequate Assessment of Need
Failing to Examine Need
Failing to Examine the Context of Need
Failing to Relate Need to Implementation Plans
Failing to Deal with Ignorance of Need
Using Needs Assessments in Program Planning
Summary and Preview
Study Questions
Additional Resource

7 Monitoring the Implementation and the Operation of Programs

The Structure of this Chapter
Introduction to Monitoring
Monitoring as a Means of Evaluating Programs
What to Summarize with Information Systems
Relevant Information
Actual State of Program
Program Participants
Providers of Services
Program Records and Information Systems
Problems with Agency Records
Increasing the Usefulness of Records
Avoiding Common Problems in Implementing an Information System
Guard against the Misuse of the Information
Avoid Setting Arbitrary Standards
Avoid Serving the Needs of Only One Group
Avoid Duplicating Records
Avoid Adding to the Work of the Staff
Avoid Obsession with Technology
Summary and Preview
Study Questions
Additional Resource

8 Qualitative Evaluation Methods

Evaluation Settings Best Served by Qualitative Evaluations
An After-School Program
Evaluating a Political Campaign
Gathering Qualitative Information
The Central Importance of the Observer
Observational Methods

Interviewing to Obtain Qualitative Information
Carrying out Qualitative Evaluations
Phase I: Making Unrestricted Observations
Phase II: Integrating Impressions
Phase III: Sharing Interpretations
Phase IV: Preparing Reports
Are Qualitative Evaluations Subjective?
Coordinating Qualitative and Quantitative Methods
Bringing Numbers to Life
Getting Insights from the Most Successful Participants
Gaining Insights from What Works Well
Changing Emphases as Understanding Expands
The Evaluation Questions
Cost of Evaluation
Philosophical Assumptions
Summary and Preview
Study Questions
Additional Resource

9 Outcome Evaluations with One Group

One-Group Evaluation Designs
Observe Only after the Program
Observe before and after the Program
Uses of One-Group, Descriptive Designs
Did the Participants Meet a Criterion?
Did the Participants Improve?
Did the Participants Improve Enough?
Relating Change to Amount of Service and Participant Characteristics
Threats to Internal Validity
Actual Changes in the Participants Not Related to Program Participation
Apparent Changes Dependent on Who Was Observed
Changes Related to Methods of Obtaining Observations
Effects of Interactions of These Threats
Internal Validity Threats Are Double-Edged Swords
Construct Validity in Pretest–Posttest Designs
Overinterpreting the Results of One-Group Designs
Usefulness of One-Group Designs as Initial Approaches to Program Evaluation
Assessing the Usefulness of Further Evaluations
Correlating Improvement with Other Variables
Preparing the Facility for Further Evaluation
Summary and Preview
Study Questions
Additional Resource

0 Quasi-Experimental Approaches to Outcome Evaluation

Increasing the Number of Times Observations Are Made
Patterns of Outcomes over Time Periods
Analysis of Time-Series Designs
Observing Other Groups

Nonequivalent Control Group Designs
Problems in Selecting Appropriate Comparison Groups
Regression Discontinuity Design
Observing Other Dependent Variables
Combining Designs to Increase Internal Validity
Time-Series and Nonequivalent Control Groups
Selective Control Design
Summary and Preview
Study Questions
Additional Resource

1 Using Experiments to Evaluate Programs

Experiments in Program Evaluation
Benefits of Experiments
Experimental Designs
Objections to Experimentation
Don't Experiment on Me!
We Already Know What Is Best
I Know What Is Best for My Client
Experiments Are Just Too Much Trouble
The Most Desirable Times to Conduct Experiments
When a New Program Is Introduced
When Stakes Are High
When There Is Controversy about Program Effectiveness
When Policy Change Is Desired
When Demand Is High
Successfully Implementing and Interpreting an Experimental Design
Take Precautions before Data Collection
Keep Track of Randomization While an Experiment Is in Progress
Analyze the Data Reflectively
Summary and Preview
Study Questions
Additional Resource

2 Analyses of Costs and Outcomes

Program Costs
Types of Costs
An Example Budget
The Necessity of Examining Costs
Comparing Outcomes to Costs
An Illustration of Cost–Benefit Analysis
The Essence of Cost-Effectiveness Analysis
When Outcomes Cannot Be Put into the Same Units
Some Details of Cost Analyses
Units of Analysis Must Reflect the Purpose of the Program
Future Costs and Benefits Are Estimated
Who Pays the Costs and Who Reaps the Benefits?
Using Cost–Benefit and Cost-Effectiveness Analyses
Major Criticisms of Cost Analyses

The Worth of Psychological Benefits Is Hard to Estimate
Placing a Value on Lives Seems Wrong
Cost–Benefit and Cost-Effectiveness Analyses Require Many Assumptions
Summary and Preview
Study Questions
Additional Resource

3 Evaluation Reports: Interpreting and Communicating Findings

Developing a Communication Plan
Explore Stakeholder Information Needs
Plan Reporting Meetings
Set a Communication Schedule
Personal Presentations of Findings
Need for Personal Presentations
Content of Personal Presentations
Audience for the Personal Presentations
Distributing Drafts of Reports
Content of Formal Written Evaluation Reports
Remember the Purposes of the Formal Report
Provide an Outline and a Summary
Describe the Context of the Evaluation
Describe the Program Participants
Justify the Criteria Selected
Describe the Data-Gathering Procedures
Provide the Findings
Develop Recommendations
Formal Reports Should Look Attractive
Provide Progress Reports and Press Releases
Summary and Preview
Study Questions
Additional Resource

4 How to Encourage Utilization

Obstacles to Effective Utilization
Constraints on Managers
Value Conflicts among Stakeholders
Misapplied Methodology
Evaluating a Program at Arm's Length
Ignoring What Is Going Really Well
Dealing with Mixed Findings
Don't Abdicate Your Responsibility
Don't Take the Easy Way Out
Show How to Use the Evaluation to Improve the Program
Using Evaluations When an Innovative Program Seems No Better Than Other Treatments
When Can Evaluators Be Sure Groups Do Not Differ?
Are Evaluations Valuable Even When No Advantage for the Innovation Is Found?
Developing a Learning Culture
Work with Stakeholders
Adopt Developmental Interpretations

Frame Findings in Terms of Improvements
Treat Findings as Tentative Indicators, Not Final Answers
Recognize Service Providers' Needs
Keep Evaluation Findings on the Agency's Agenda
The Evaluation Attitude
Humility Won't Hurt
Impatience May Lead to Disappointment
Recognize the Importance of the Evaluator's Perspective
Work on Practical Questions
Work on Feasible Issues
Avoid Data Addiction
Make Reports Accessible
Seek Evaluations of Your Own Work
Encourage the Development of a Learning Culture
Summary and Possible Trends for Program Evaluation
Study Questions
Additional Resource

References
Author Index
Subject Index

Preface

Welcome to the exciting field of program evaluation and to this introduction to the practice of it. If you do not already know that program evaluation is a fascinating area, I encourage you to pay particular attention to the Evaluator Profiles throughout this book and the accounts of how many of them discovered this almost hidden treasure. As you will learn, it involves a set of skills and practices that can provide valuable insights to crucial human services as well as provide important components for careers, or even an entire career. But unlike many occupations such as medical doctors or professors, too many people know little or nothing about it, so few children say, "I want to be a program evaluator when I grow up." Many current evaluators can tell wonderful stories about how they discovered this great field that they never knew existed.

This textbook is intended as a solid introduction to the field for those who are ready to build a good foundation. On the one hand, it covers all the main aspects and provides substantial practical insights. Students who learn this material diligently, especially if they also have the opportunity to have some hands-on experiences while doing so, should be prepared then to venture into formal program evaluations, although many will be glad to start as part of a team. On the other hand, there are a number of specific aspects of professional program evaluation that would benefit from further specialized study, such as particular approaches to evaluation like empowerment and utilization, or specific forms of evaluation like needs assessment. This good introduction does not pretend to be fully comprehensive, and those who are ready to tackle the full field have many excellent advanced books and formal programs to choose from.

Still, this ninth edition builds on a fine tradition even as it ventures into some new territory. Emil Posavac's editions have made a long-standing contribution to the field. I (Ken Linfield) have used his textbook every year since I started teaching our doctoral course now twelve years ago, and it was the one used for years before I started. As I join in writing this new edition as the first author, I am especially pleased to expand the case studies and profiles to all fourteen chapters, providing more extended examples of a broader range of evaluations and evaluators. The case studies provide details from the evaluators' perspectives even as they have also been organized with the MISSION framework introduced in this edition. The MISSION acronym is presented to help evaluators, especially those newer to the field, to remember easily the broad range of elements in program evaluation—Meeting needs, Implementation, Stakeholders, Side effects, Improvement focus, Outcomes, and Nuances (see especially Chapter 1).

Another extension of the solid material in this text matches the developments of our time, adding resources available online to the hard copy textbooks or the e-books that present the standard chapters. These "eResources" provide an important span of information that would not fit in a traditional textbook—multiple examples of evaluation reports, extensive descriptions and summary tables of the growing number of models of evaluation, and much more. Further, active hyperlinks and other materials like spreadsheets provide a dynamic option impossible with traditional textbooks. And not only do the eResources also permit some level of interaction with the first author, there are some lighthearted sections as well—check out the video of my "Program Evaluator" song and others. Perhaps more importantly, regular updates will address changes since the publication of the book you are reading.

So—welcome to the great field of program evaluation. I invite you to find the portions that particularly connect with what you are passionate about—improving services for people who have endured traumatic events, attending to a population that is important to you, building on essential values like evidence and integrity, or another. Enjoy the ride!

Ken Linfield
February 2018

1 Program Evaluation
An Overview

Suppose you are working for the Student Counseling Center at Western Tech in a year or two. The center plans to offer a Sexual Assault Prevention Program to women and men in the following month, and you have been asked to develop a plan that will permit the center's staff to decide whether (a) to offer the program again because it is successful, (b) to alter the program in order to make it more useful to participants, or (c) to drop the program because it fails to meet a need. You might stop reading for a moment and write down the steps you think are necessary to measure the success of such a program. Among other points, consider what evidence would indicate the program is successful, what would mean that it should be improved, and what would show that it should be ended.

Because the center wishes to maintain and promote useful services to students, evaluating a program that might reduce sexual assaults on campus is an important assignment. The staff of the center would not want to devote resources for a program that does not meet the needs of students on campus; in such a case, something more helpful should be offered instead. If the program is good, you would want your plan to detect its strengths; where it needs improvement, you would want to detect its limitations. When you are finished with this text, you will be ready to tackle this project. To begin, what is program evaluation?

The practice of evaluating one's own efforts is as natural as breathing. Cooks taste their gravy and sauce and basketball players watch to see whether their shots go in. Indeed, it would be most unwise after turning on the hot water to neglect to check the water temperature before stepping into a shower. At a basic level, program evaluation is nothing more than applying this commonsense practice to settings in which organized efforts, called "programs," are devoted to helping people who need education, medical treatment, job training, safe streets, welfare assistance, safety while traveling in airplanes, recreational services, or any of the thousands of services provided in a modern society. Program evaluation is a methodology to learn the depth and extent of need for a human service and whether the service is likely to be used, whether the service is sufficient to meet the unmet needs identified, and the degree to which the service is offered as planned and actually does help people in need at a reasonable cost without unacceptable side effects. Utilizing research methods and concepts from psychology, sociology, administration and policy sciences, economics, and education, program evaluators seek to contribute to the improvement of programs.

There are several crucial differences between the natural, almost automatic, evaluations we carry on as we work and play and the practice of program evaluation in organizational settings. These differences make program evaluation harder than the self-evaluations that we all do. First, organizational efforts are nearly always carried out by a team—whether that be a school team of teachers, coaches, and administrators; a hospital team of physicians, nurses, and technicians; or a welfare department team of social workers, counselors, and file clerks. This specialization means that the responsibility for program effectiveness is shared among many people. Furthermore, the results of such programs—an educated student, a well patient, or an effective young mother—are not the sole responsibility of any individual.

Second, most programs attempt to achieve objectives that can only be observed sometime in the future rather than in a matter of minutes, as is the case with a cook adding a spice to a simmering pot or a painter repairing a damaged wall. In other words, as the time between an activity and the desired outcome of that activity lengthens, it becomes less clear what we are to observe in order to decide that the activity is being carried out appropriately or not and what could be done to improve the result of the activity. In fact, the choice of criteria to use in deciding whether a program is working well creates debates among evaluators, program personnel, clients, and funding agencies. How to respond when a problem is detected is often just as controversial.

Third, in contrast to evaluating our own efforts, there are many different parties involved in the evaluations of programs. For example, the effectiveness of teaching in a school may be assessed by evaluators who work for the central administration or the school board. It would not be surprising that individual staff members might feel threatened if the evaluators do not have their trust because, in an extreme case, the livelihood of the teachers might be harmed by an incompetently done program evaluation.

Last, programs are usually paid for by parties other than the clients of a program. Funds for the salaries of nurses are obtained from payments to hospitals by insurance companies and government agencies, and schoolteachers are paid through taxes collected by school districts. Although nearly all nurses and teachers want to do effective work, their job security is at least in the short term more dependent on keeping the program funders satisfied than serving their patients and students well.

Such complications as these make program evaluation considerably more difficult to do than the evaluations of our own activities that we perform so easily and naturally. This text focuses its discussion on evaluations of programs in human services, such as education, medical care, welfare, rehabilitation, or job training. Program evaluation is one form of the general discipline of evaluation that includes employee assessment, quality control in manufacturing, and policy analysis, among other activities (Scriven, 2003). The various kinds of evaluation are sometimes confused with each other. All forms of evaluation involve discovering the value and worth of something; however, the focuses, purposes, and methodologies are quite different.

EVALUATION AREAS THAT NEED TO BE CONSIDERED

A common assumption by people who are not familiar with program evaluation is that all evaluators do is measure the outcome of a program—perhaps counting how many of the clients in a substance abuse treatment program no longer use the substance, either at the end of the program or at a later time, or determining the typical change in knowledge (hopefully an increase) for those who have attended a series of workshops. As will be seen below, measuring outcomes is a very important part of program evaluation. But it is only one part. We will make the case briefly here and then repeatedly throughout this text that program evaluation is a much broader set of activities than just measuring the outcome of a program and that good evaluators consistently think about this wider range of elements. The following paragraphs introduce the seven general areas that program evaluators need to consider. It should be noted that some evaluations may not explicitly include all these elements, and at times a single one of the components may be a complete evaluation (for example, a needs assessment). But choosing not to include one or more elements should be a deliberate decision based on a clear rationale, not a matter of focusing so intently on a limited set of areas that one forgets other considerations.

Meeting Needs

Most often, program evaluations are conducted on existing programs where there is a reasonable likelihood that the programs will continue in somewhat similar fashion for the foreseeable future. As a result, the original need that led to the program may not be raised as an important point of the evaluation; nearly everyone involved may just assume that doing something to address that need is a worthwhile activity. In fact, many people may be unable to imagine not having certain programs. Most of us assume that education is necessary for children and that we must have police and fire departments in our communities without needing to document precise details of what those children do not yet know or of the potential for crime and fires. As you will see in Chapter 6, "The Assessment of Need," there are a number of important aspects of this component of program evaluation beyond just documenting that there is a legitimate problem that the program addresses.

One overarching point, however, is a focus on unmet needs. Although we all need to breathe air, we do not have large numbers of programs providing air nor evaluations to discuss them. Rather obviously, programs are developed when one or more needs are not already addressed. So evaluators consider unmet needs—whether in an area that has never been addressed or the extent to which current needs are not fully met by existing programs. Various other aspects can also be considered, such as how the need might be expected to change in the future or how the need is perceived differently by different people.

Implementation

A common problem with programs is that some are either never implemented as planned or are implemented in such a way that people in need receive no or minimal benefit. Consequently, it is essential that evaluators confirm points ranging from the fact that program personnel have been hired and offices have been opened to documenting how many clients have found the program and the nature of the services provided. When the primary focus addresses such issues, the project may be called an implementation evaluation, in contrast to an outcome evaluation that also considers the results of the interventions.

Although the monitoring function of program evaluation may at first strike the reader as unnecessary, there are many examples of programs that, although funded, were never put into practice. Even more common, descriptions of programs in agency materials may be accurate only with regard to what was intended, but not what is currently happening. In fact, a recent estimate notes "that failure rates for organizational change implementation typically range from 28% to as high as 93%" (Durand, Decker, & Kirkman, 2014, p. 404). To be fair, the actual practices could be better than originally planned, rather than worse. In any case, an essential area for evaluators is monitoring the details of how the program has been implemented.

Stakeholders

In the process of conducting an evaluation, evaluators might generally expect to have some contact with the clients who are served by the program, especially indirect contact, by considering data from the clients such as information on their changed condition or their responses to a questionnaire. But it is essential that evaluators take into account the many people who are involved with a program in one way or another, those who provide the service as well as those who receive it, and those who are supportive from a distance as well as those whose connection is either mixed or even opposed to some degree. The idea that all of these people have a stake of some form in the program has led to the term "stakeholders" as a general way to talk about them.

Staff providing the services are obviously highly involved both because they deal day-to-day with the clients and because they generally are paid to provide the services. Paying attention to staff as stakeholders involves many points such as recognizing that therapists often know important details about clients beyond what is contained in agency records, that staff may either resist change or be eager for it, and that most employees have some level of apprehension about the results of an evaluation in terms of their own future. Of course, there are a range of staff positions in most agencies—therapists, administrators, support staff, and others. In addition, most agencies have a board that oversees general operations. Often there are others in the community who care about the program in some way—typically including family and perhaps friends of clients, people in the community who are passionate about the specific need being addressed, and occasionally those who are affected by financial aspects of the program. At times, those who live near a treatment facility may have reactions to details of the operation such as traffic or clients' behavior or still other points.

On the one hand, it is impossible for evaluators to know everything about all stakeholders and completely unrealistic to try to make all stakeholders happy. But on the other hand, it is essential that evaluators make reasonable efforts to learn about the major stakeholders and any important current issues. A common element is that the staff providing services, administrative staff, and board members have different expectations for the evaluation itself. If administrators want to improve efficiency, board members want to lower costs, and therapists want more freedom to extend treatment, there are important implications for the progress of an evaluation and what evaluators should expect regarding cooperation from the stakeholders.

Recently, evaluation researchers have explored the principles that appear to support successful collaboration with stakeholders, such as clarifying the motivation for collaboration (Shulha, Whitmore, Cousins, Gilbert, & al Hudib, 2016) as well as what can be learned about the involvement of stakeholders in evaluation more generally, such as noting that studies have found substantial evidence both for benefits and for liabilities (Brandon & Fukunaga, 2014).

Side Effects

Efforts to solve problems usually create some undesirable side effects (Sieber, 1981). For example, medications

that relieve one health problem may create problems in other systems of the body; sometimes these side effects are minor, sometimes they require medical attention or a change in medication, and sometimes, though rarely, they are fatal. Patients and physicians must therefore weigh the dangers of side effects, neither minimizing their importance nor avoiding helpful treatments. Even when the benefits appear clearly to outweigh the costs to medical professionals, others may perceive these effects differently. An important current challenge is that some parents come to different conclusions about the comparative value of vaccinations than physicians (Harmsen, Mollema, Ruiter, Paulussen, de Melker, & Kok, 2013). A parallel situation occurs in the area of educational and social service programs. Welfare policies can help people during crises to reestablish their productive roles in society; on the other hand, welfare policies can create long-term dependency (McKnight, 1995). Programs to provide social support to isolated people can have negative effects that have typically been ignored (Lincoln, 2000). Special education classes permit children with unique needs to learn at a comfortable pace, yet being in such a class carries a stigma. Consequently, program planners seek to develop services that provide benefits but minimize negative side effects. On the other hand, unforeseen side effects are occasionally positive: When people learn new skills, self-esteem often increases (Dawes, 1994; Felson, 1984).

Improvement Focus

Most of these seven areas cover elements of programs that good evaluators consider, measure, and even analyze with formal statistical procedures. The area of improvement focus, in contrast, describes an attitude or approach that evaluators take throughout the course of an evaluation. This pervasive aspect involves considering not just how good the program is now, but how it can become better in the future. The improvement focus is especially important when the evaluation occurs early in the life of a program and is intended to catch and fix any problems. Such evaluations are often called formative evaluations in contrast to summative evaluations that may come at or even after the conclusion of the program. Formative evaluations explicitly look for ways to improve the program, whether by correcting actual mistakes or unanticipated problems or by building on particularly successful components. Even with summative evaluations, however, an improvement focus will make an important contribution whether the program is directly replicated or adapted to fit new settings, circumstances, or methods.

To be clear, this improvement focus does not mean that evaluators are never critical of problems or shortcomings, let alone of failures in a program. Remember that the primary purpose of an evaluation is to determine the value or worth of a program, so if a particular intervention is wasting time or money, or, worse, actively harming clients, evaluators have an obligation to report on these problems clearly. But even with substantial limitations, most programs have value that can be identified and used as a basis for better work in the future. Although it is possible that an evaluation may find there is nothing redeemable in a given program, such conclusions are rare especially when evaluators take the improvement focus seriously.

Outcomes

It should be fairly obvious that an evaluation of a program would consider the results such as how clients improve or other details about how those involved in the program have changed because of their participation. If the proof of the pudding is in the eating, the proof of the program is commonly seen in those who have gone through the program. Those who attend educational programs are expected to know more. Clients in treatment for problems should show fewer (or no) problems; they might also show higher levels of functioning —increased strengths, not just fewer weaknesses. Broader programs such as in public health might be expected to lead to changes at the community level—lower incidence of particular diseases or other problems among the entire population, not just with identified individual. The concept of outcomes is a general one, but the details vary considerably depending on the particular program.

As noted above, for many people examining outcomes is the most obvious, if not the only, aspect they look for in an evaluation. Although it is essential that good evaluators think about the other areas, outcomes need substantial attention, too. Improvement in clients could be seen in a reduction of problems, in an increase of skills, or in both. Clients' own perception of skill may be important as may observations by others that their abilities have increased. The most appropriate time frame for documenting changes might be at the end of treatment, six months after treatment, or six years later. Increasingly, it is not enough for evaluators to show

that those who participated in programs are different following the program—there is an emphasis on showing that the program contained the "active ingredients" that caused the change (Gates & Dyson, 2017). Chapters 8 through 11 will address the many ways that evaluations consider evidence, not just that there are good outcomes, but that the programs are responsible for those good outcomes. As you will see, at times there are at best gentle hints that the program may have led to the changes whereas under the best conditions, there may be very clear reasons to believe that the program caused the good effects. Evaluators examine outcomes and the evidence that those outcomes are caused by the program.

When the specifics about outcomes are clear, the task may still not be finished. There are a number of ways to compare aspects of the outcomes that are important in some evaluations. Although many agencies have one program with a single focus and the goal is simply to have good results from it, others have multiple programs with more sophisticated criteria and where programs are compared with one another in terms of the value of the benefit overall or in terms of the cost for a given benefit. A program with moderately good outcomes may not have good enough results to justify its continuation if resources are threatened. The larger area of outcomes can thus be subdivided into various kinds of comparisons—by cost, by value to the stakeholders, or by other constraints such as how long it takes to achieve a given result. All these ways of examining outcomes are potential areas for evaluators to consider, and will be addressed in a variety of ways later in this text.

Nuances (Mechanisms)

Although evaluations have often aimed simply to document that participants have benefitted sufficiently, another growing trend is to examine more sophisticated questions such as specific reasons why the good outcomes have occurred—both in terms of what mechanisms were involved as well as considering how those in different groups or conditions experience better or worse results (Solmeyer & Constance, 2015). Were all the services provided (and funded) essential ingredients in the good results or were some of the activities pleasant for the clients but otherwise irrelevant ways to use staff time, effort, and the funders' money? Elements of therapy that continue out of habit or tradition rather than because they have been shown to lead to improvements are good candidates to be reconsidered and, potentially, eliminated. Another important point is that some of those receiving services may not benefit like the rest of the group either because of their existing characteristics or because they react to the treatment differently. A treatment that works well with older adolescents but is not effective with the youngest clients might appear to yield good outcomes overall, but understanding the different effects could lead to using a different, but better, treatment for the younger children.

One important aspect of examining mechanisms in an evaluation is that measures of the suspected mechanisms must be included. If therapists or evaluators believe that cognitive restructuring is an essential part of useful therapy, the evaluators will need to assess how much each client successfully restructured thoughts. With such measures, the evaluators can analyze the degree to which clients with greater restructuring tended to improve more. If the effect is consistent and large enough, that would provide evidence to support the use of that component. In addition to specific techniques, a somewhat broader construct such as a measure of the degree to which the planned intervention is followed can also address the question of mechanisms (Abry, Hulleman, & Rimm-Kaufman, 2015). More detailed discussions of these specifics will appear in the sections on data analysis, especially Chapter 11 on experimental approaches.

MISSION

Many evaluators new to the profession find it challenging to remember these aspects of good evaluations. To help students and others, "MISSION" is an acronym for these seven areas—Meeting Needs, Implementation, Stakeholders, Side Effects, Improvement, Outcomes, and Nuances. These seven elements provide an all-purpose set of categories for evaluators to consider regardless of the evaluation. As noted above, it is possible that in a particular case, the evaluator might choose to minimize or even ignore the issue of needs if all stakeholders were already so thoroughly convinced of their reality. But not addressing one of these areas should be a deliberate choice. And more than one evaluator has forgotten to pay attention to side effects only to regret the oversight later. To help students remember these elements, the MISSION acronym will be a

theme throughout this text, explicitly used in the case studies to illustrate both the various forms these components can take as well as the times when good evaluations may pay little or even no attention to some of them.

COMMON TYPES OF PROGRAM EVALUATIONS

The primary goals of program evaluation, just discussed, can be met using a number of different types of program evaluations; the major ones involve studies of need, process, outcome, and efficiency.

Assess Needs of the Program Participants

An evaluation of need seeks to identify and measure the level of unmet needs within an organization or community. Assessing unmet needs is a basic first step before any effective program planning can begin. Program planning involves the consideration of a variety of alternative approaches to meet needs. In selecting some alternatives and discarding others, planners engage in a form of program evaluation, one that occurs before the program is even begun. The close association between program planning and program evaluation is indicated in the title of the journal *Program Planning and Evaluation*.

As part of the assessment of need, evaluators may examine the socioeconomic profile of the community, the level of social problems within the community, and the agencies and institutions currently serving the community. Through close contact with residents and local leaders, evaluators can learn which aspects of a program are likely to be useful and which might be unacceptable, thus adding depth to the inferences drawn from statistical summaries suggesting critical unmet needs. Sometimes program evaluators believe that there is a pressing need for a sense of self-determination or empowerment (Fetterman & Wandersman, 2004).

Examine the Process of Meeting the Needs

Once a program has been developed and begun, evaluators turn to the task of documenting the extent to which implementation has taken place, the nature of the people being served, and the degree to which the program operates as expected. The activities making up the program, such as hours of counseling, number of classes held, and hours police officers patrolled, are called "outputs." Evaluations of process involve checking on the assumptions made while the program was being planned. Do the needs of the people served or the community match what was believed during planning? Is there evidence to support the assessment of needs made during the planning stage? Do the activities carried out by the staff match the plans for the program? What evidence can be found that supports the theoretical assumptions made by the program planners? It is crucial to learn how a program actually operates before offering the same services at additional locations or with other populations.

Under normal situations, the information necessary for an evaluation of process is available in the records of an agency sponsoring the program; however, the information may be recorded in a way that is hard to use. For example, information on application forms often is not summarized, and records of actual services received may not permit easy analysis. Furthermore, sometimes no records are kept on a day-to-day basis. In Chapter 7 on program monitoring and information systems, some very simple methods of developing an information system for a human service setting are illustrated. In addition to quantitative information, an evaluation of process may well benefit from unstructured interviews with people using and not using the service. Such qualitative information often provides points of view that neither evaluators nor service providers had considered.

Measure the Outcomes and Impacts of a Program

If a study of implementation shows that a program has been implemented well and that people seek and secure its services, a measurement of the program's outcomes may become a focus of an evaluation. An evaluation of outcome can take on several levels of complexity. The most elementary level concerns the condition of those who have received services. Are program participants doing well? Do they possess the skills being taught? A more challenging evaluation would compare the performance of those in the program with

those not receiving its services. An even more challenging evaluation would show that receiving the program's services caused a change for the better in the condition of those participating in the program (Boruch, 1997).

Although program managers hope that their programs will elicit positive changes in people, the causes of behavioral changes are difficult to pin down. For example, many people begin psychotherapy during a life crisis. If after several months of counseling they feel better, the change could be due to the counseling, the natural resolution of the crisis, or something else entirely. In a work setting, a change in procedures may result in better morale because of increased efficiency or because the workers feel that management cares about their well-being. Or perhaps an improved national economic outlook reduced the workers' anxiety about possible job loss. Discovering the causes of behavioral changes requires sophisticated techniques since an organization must continue to provide services while the evaluation is being conducted. Experienced evaluators are not surprised by tensions between gathering information and providing services.

Besides the limitations on the choice of research design, evaluators seeking to assess the outcome of a program often discover that people hold different opinions about what constitutes a successful outcome. A job-training program paying unemployed people to learn job skills may have been planned so that they can later obtain jobs with private companies. City officials may view the training as a reward for people who have worked in local election campaigns, and the trainees may view the program as a good, albeit temporary, job. An evaluator would not adopt just one of these views of the desired outcome of the program.

Assessing the maintenance of improvement creates another problem when evaluating outcomes. People leaving a drug rehabilitation program typically return to the same community that contributed to their problems in the first place. Despite their good intentions, people treated for alcoholism are often unable to withstand the influence of their peers. Changing long-standing behaviors is difficult. Although positive changes may be observed after a person's participation in a program, the changes may be only superficial and disappear in a matter of months, weeks, or days. If so, was the program a failure? When positive outcomes do occur, their effects can influence other behaviors and even improve the condition of other people (e.g., children); such long-term outcomes are called "impacts."

Integrate the Needs, Costs, and Outcomes

Even when evaluators can show that a program has helped participants, they must also deal with the question of costs. Just as a family must make choices in how a budget is spent, governments and agencies must select from among those services that might be supported. A successful program that requires a large amount of resources simply may not be a good choice if a similar outcome can be achieved with lower cost. If an evaluator is asked to compare two or more programs designed to effect similar outcomes, efficiency can be assessed in a fairly straightforward manner; this would be called a cost-effectiveness analysis (Levin & McEwan, 2001). Real difficulties occur when evaluators are asked to compare the efficiency of programs designed to achieve different outcomes, sometimes for different groups of people. For example, should a university spend funds to reduce alcohol abuse among students or to increase the number of tutors available? Although competing purposes are seldom contrasted in such a stark fashion, no organization can do everything that might be worthwhile; choices have to be made. Evaluations that incorporate high level integration of multiple elements have the potential to provide information to administrators to help them make such choices (King, 2017).

Note that there is a logical sequence to these four general types of evaluations. Without measuring need, programs cannot be planned properly; without a plan, implementation would be haphazard; without effective implementation, outputs are not provided adequately and good outcomes cannot be expected; and without achieving good outcomes, there is no reason to worry about efficiency. A premature focus on an inappropriate evaluation question is likely to produce an evaluation with little value (Wholey, 2004).

IMPORTANT ISSUES IN EVALUATIONS

Time Frames of Needs

Short-Term Needs

Many programs are set up to help people in crises. For example, medical care is needed for injuries or illnesses, and emotional and financial support are often needed after a crime, a natural disaster, or a home fire. Some educational services are designed to meet specific needs of employees required to learn new skills. To be effective, such services should be available on short notice.

Long-Term Needs

To meet some needs, services must be available for a long period of time. Children need education for many years. Psychotherapy, treatment for chronic illnesses, training for prison inmates, and rehabilitation for alcoholism and drug abuse are examples of programs that focus on problems that cannot be resolved in a short time period.

Potential Needs

Ideally, potential problems can be averted. Preventive programs are supported so that problems can be avoided or at least postponed. The effectiveness of immunization, health education, or business security programs is judged by the degree to which problems do not occur. Clearly, evaluations of prevention programs must differ from evaluations of programs developed to relieve current problems.

Extensiveness of Programs

Some programs are offered to small groups of people with similar needs, others are developed for use at many sites throughout a county or a state, and yet others are written into federal laws to be offered throughout a nation. Although tools useful in program evaluation apply to evaluations carried out at all levels, there are considerable differences between an evaluation of a day hospital program in Memorial Hospital and an evaluation of psychiatric services supported by Medicaid. Although it would be quite reasonable for these evaluations to use some of the same measures of adjustment, the decisions faced by the administrators of the local program differ so greatly from those faced by Medicaid managers that the evaluations must differ in scale, focus, and type of recommendations.

The discussions throughout this text focus on evaluations of smaller programs likely to be encountered by evaluators working for school districts, hospitals, personnel departments, social service agencies, city or state governments, and so on. Although some national programs are mentioned, most readers of an introductory text are more familiar with local government or educational programs than they are with national programs. Even when a program is supported by federal funds, implementation at a local agency differs markedly from those in other localities. The characteristics of programs are not like medications, which can be given in easily specified 20 mg doses. Thus, organizations with support from local governments, foundations, and federal agencies are all expected to carry out program evaluations and submit them to the funding source.

For studies done at a local level, statistical procedures seldom need to be particularly complicated partially because many administrators find statistics to be an incomprehensible—if not intimidating—topic. Complicated analyses require data from larger numbers of people than are usually available to evaluators working with local organizations. Evaluators can often show how to improve data summaries and presentations so that program information can be used to answer questions faced by program administrators.

PURPOSE OF PROGRAM EVALUATION

There is only one overall purpose for program evaluation activities: contributing to the provision of quality services to people in need. Program evaluation contributes to quality services by providing feedback from program activities and outcomes to those who can make changes in programs or who decide which services are to be offered. Without feedback, human service programs (indeed, any activity) cannot be carried out effectively. Our bodily processes require feedback systems; similarly, feedback on behavior in organizations is also crucial for the success of an organization. Delayed feedback, not clearly associated with the behavior being examined, is not very informative.

Figure 1.1 illustrates the place of program evaluation as a feedback loop for a human service program. Assessing needs, measuring the implementation of programs to meet those needs, evaluating the achievement of carefully formed goals and objectives, and comparing the level of outcome with the costs involved relative to similar programs serve to provide information from which to develop program improvements (Schermerhorn, Hunt, & Hunt, 2008; Wholey, 1991) or to make wise choices among possible programs (Levin & McEwan, 2001).

Feedback can be provided for different purposes. First, as noted earlier, evaluations can strengthen the plans for services and their delivery in order to improve the outcomes of programs or to increase the efficiency of programs; these formative evaluations (Scriven, 1967) are designed to help form the programs themselves. Alternately, summative evaluations can help us decide whether a program should be started, continued, or chosen from two or more alternatives (Scriven, 1967). There is a finality to summative evaluations; once the value and worth of the program have been assessed, the program might be discontinued. In actuality, very few single evaluations determine whether or not a program will be continued (Cook, Leviton, & Shadish, 1985). Many sources of information about programs are available to administrators, legislators, and community leaders. Because program evaluation is only one of these sources, evaluators are not surprised when clear evaluation findings do not lead to specific decisions based on the evaluation. Chelimsky (1997) uses the terms *evaluation for development* and *evaluation for accountability* to refer to the formative and summative purposes, respectively.

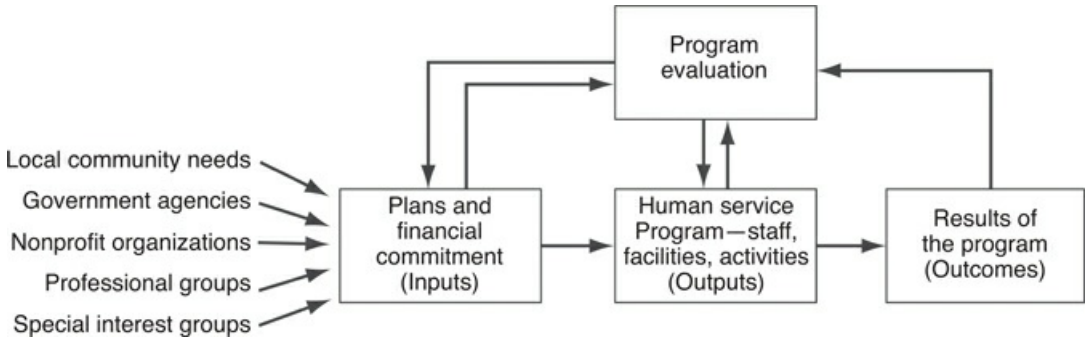


Figure 1.1 The place of evaluation as a feedback loop for a human service program.

Once a program has been carefully refined using feedback from formative evaluations and found to be effective using a summative evaluation, frequent feedback is still needed to maintain the quality of the program. This third form of evaluation is called "monitoring"; its purpose is quality assurance (Schwartz & Mayne, 2005; Wholey, 1999). This can be as simple as checking the speedometer from time to time while on a highway and as complicated as examining how international aid is used (Svensson, 1997). Monitoring critical process and outcome variables to verify that an effective program stays effective is a crucial activity after programs have been successfully implemented. Furthermore, monitoring can be expected to identify problems when the community or economic conditions change. As efforts are made to resolve those problems, monitoring feedback takes on a formative function as well. If problems cannot be resolved, monitoring can come to serve a summative purpose also.

Some evaluations are also carried out to learn about programs; Chelimsky (1997) uses the term *evaluation for knowledge* to describe these. This purpose is seldom part of the work of evaluators working on specific programs. Their work might be used by others to deepen our understanding of, for example, rehabilitation programs, but this occurs only after evaluations are completed and shared. Evaluators recognize the need to develop better understandings of programs and intervention strategies; however, this text does not focus on such types of evaluation because they involve more ambitious undertakings than most evaluators deal with regularly.

THE ROLES OF EVALUATORS

Principles, Competencies, and Credentialing

Along with the above points about components, types, and issues in program evaluations, there are additional ways to think about what makes for good evaluations and good evaluators. Although on one level, describing such expectations might seem straightforward, in fact, there is no simple answer to either of these two similar but not identical points. (Although good evaluators should be expected to produce good evaluations, assessing people and reports involve quite different processes.) Still there are a number of relevant details about the current state of the field in these areas, with both similarities and differences in approaches across the world. The main categories are Standards or Principles, Competencies, and Credentialing, and the following brief summary describes the state in fall 2017. For updates since this edition was published, see the eResources.

The American Evaluation Association (2017) has adopted "Guiding Principles for Evaluators" that cover five basic areas such as Integrity/Honesty, while the Canadian Evaluation Society (CES) has adopted standards that cover five perspectives such as Utility and Accuracy (Yarbrough, Shulha, Hopson, & Caruthers, 2011). Like many organizational principles, many of the points are more aspirational than precise. For example, one of the specific AEA principles is that evaluators should "seek to maintain and improve their competencies in order to provide the highest level of performance in their evaluations" (American Evaluation Association, 2017, p. 313), which does not indicate how much improvement is enough. Still, the intent of the principles and standards is obvious, and blatant failures to follow them can be noted.

In some areas, the general agreement in the field has not yet crystallized into uniform policies. Proposals for establishing a list of competencies are more than a decade old (King, Stevahn, Ghere, & Minnema, 2001; Stevahn, King, Ghere & Minnema, 2005) and current research in the field refers to these proposals favorably (e.g., Galport & Azzam, 2017). The details fit well with both the principles and common practice, noting crucial elements such as showing ethical approaches like respecting clients and considering public welfare, demonstrating knowledge about appropriate research methods, and being responsive to the specific context of a given program (Stevahn et al., 2005). The Canadian Evaluation Society (2010) formally adopted a set of Competencies, while the American Evaluation Association has not done so at this time, although many evaluators continue to demonstrate their voluntary commitment to those values.

The CES also offers the designation of "Credentialed Evaluator" to CES members who apply for it and meet the criteria (Kuji-Shikatani, 2015). The possibility of a credentialing process in the United States and the related possibility of an accreditation process for programs that train evaluators continue to be raised and discussed (Altschuld & Engle, 2015; King & Stevahn, 2015; Kuji-Shikatani, 2015; LaVelle & Donaldson, 2015; McDavid & Huse, 2015; Seidling, 2015; Shackman, 2015). The trade-offs between more uniform quality on the one hand and flexibility to address substantially different needs and contexts on the other have not yet been resolved.

Comparison of Internal and External Evaluators

There are two primary ways evaluators relate to organizations needing evaluation services: (1) evaluators can work for the organization itself and evaluate a variety of programs in that setting, or (2) they can work for a research firm, a university, or a government agency to evaluate a specific program. In this book, the terms *internal evaluator* and *external evaluator* refer to these two types of evaluators, respectively. Although there are great similarities in what all evaluators do, the affiliation of an evaluator has implications for the manner in which evaluations are done.

Factors Related to Competence

In terms of knowledge about a program, internal evaluators have an advantage since they have better access to program directors and to the administrators of the organization. A person who is physically present on a routine basis is likely to see programs in action, to know staff members, and to learn about the reputation of programs from other people in the organizations. Such information is unavailable to an external evaluator. The more that is known about the actual operation of programs, the easier it is to ask relevant questions during the planning and interpretation of evaluations.

The technical expertise of an evaluator is important. An internal evaluator often works with a small group of two or three; some evaluators work alone. An external evaluator, however, can often draw upon the

resources of a greater number of people, some of whom are very skilled in sampling, qualitative analyses, or statistics. Some independent evaluators, however, essentially work alone and do not have access to such resources.

A different facet of the question of technical expertise is suggested by the need to perform evaluations of programs in different areas of an organization. Moving from topic to topic is stimulating; however, there is also a risk that inexperience with the services being evaluated could limit an internal evaluator's insight into aspects of some programs. By selecting external evaluators with experience in the type of program being evaluated, organizations might avoid errors due to inexperience.

Personal Qualities

In addition to technical competence, an evaluator's personal qualities are also important. Evaluators can do more effective work if they are objective and fair, trusted by administrators, and concerned that program improvements be encouraged. Being well known and having been found worthy of trust (Taut & Alkin, 2003), internal evaluators usually find program directors and staff more willing to devote time to an evaluation, to admit problems, and to share confidences than they would be with an evaluator not affiliated with the organization. Being trusted increases the likelihood that a competent evaluator can fill an organizational educator role. Since credibility can be easily lost, evaluators avoid acting in ways that might suggest that they have allowed personal biases to influence their conclusions.

An internal evaluator can also be expected to want to improve the organization sponsoring the program and paying the evaluator's salary. External evaluators, being less dependent on the organization sponsoring the program, might not have the same level of commitment to working for program improvements. On the other hand, being dependent on the organization could affect the objectivity of an evaluator. Knowing program directors and staff members might make it difficult for an internal evaluator to remain objective (Ray, 2006; Scriven, 1997a); it is not easy to criticize a friend. An external evaluator is less likely to experience conflicting pressures when an evaluation reveals problems in a program. Internal evaluators might find it easier to remain objective if they remember that their value to an organization depends on their work being credible. Developing a reputation for tackling sensitive issues is somewhat easier when evaluators emphasize that most deficiencies in programs are due to system problems rather than personal inadequacies. Deming (1986), an early champion of the quality revolution in industry, insisted that 85% of problems in workplaces are due to limitations in organizational procedures or work setting designs and 15% due to staff problems. Since many people mistakenly believe that evaluations focus on finding bad apples, evaluators should often refer to Deming's point.

Factors Related to the Purpose of an Evaluation

Earlier in this chapter the purposes of evaluation were discussed—formative, summative, and quality assurance. Both internal evaluators and external evaluators can perform any type of evaluations; however, internal evaluators may have an advantage in performing formative and quality assurance evaluations. This seems true because such evaluations cannot prompt a decision to end a program. In fact, the internal evaluator's rapport with managers and staff may encourage nondefensive communications, which are essential if avenues to improve programs are to be found. If, by contrast, a summative evaluation is wanted because support for a program might be withdrawn, it may be better if an external evaluator is used instead of an internal evaluator. For example, if a small elementary school district knows that it must close one of three schools, it may be wise to have an external evaluator perform an evaluation of the schools and select the one to close. No matter which one is selected, some residents will object. It could be hard for an administrator to serve in the school district if a sizable portion of the residents were dissatisfied with the decision; however, the external evaluator need never set foot in the school district again.

As should be clear, internal and external evaluators have many complementary strengths and weaknesses. Depending on the specifics of a given evaluation, one or the other may be preferable. A growing possibility in recent years is to develop a partnership between internal and external evaluators that has the potential to make good use of the strengths from both (Le Menestrel, Walahoski, & Mieke, 2014).

Evaluation and Service

An evaluator's job falls somewhere between the role of the social scientist concerned about theory, the design of research, and the analysis of data (but for the most part uninvolved with the delivery of services) and the role of the practitioner dealing with people in need (but seldom deeply interested or extensively trained in the methods of data collection and analysis). Evaluators are able to use the language and tools of the research scientist; however, they must also be sensitive to the concerns and style of the service delivery staff. In addition, evaluators are called upon to communicate with organization administrators who have different priorities, such as balancing a budget and weighing proposals for expansions of services. Since the role of program evaluator is still fairly new, it is likely that the evaluator will at times seem out of step with everyone.

Participating in such a new field has advantages and disadvantages. The advantages include the intellectual stimulation provided by exposure to people serving various roles in human service settings and the satisfaction of seeing research methods used in ways that can benefit people. The disadvantages include being viewed as intrusive and unnecessary by some service delivery personnel. Instead of concentrating on problems, sensitive evaluators make sure program managers and staff have opportunities to talk about what goes well and what is rewarding in their work (Preskill & Coghlan, 2003). Yet to demonstrate the value and worth of the program, effective evaluators must ask challenging questions and tactfully insist that answers be supported by data.

Of course, at times even the most skilled evaluator will be in conflict with an organization. One of the early leaders in evaluation was concerned that the perceived threat of evaluation would make it impossible to collect data of sufficient validity to draw useful conclusions (Campbell, 1983). By contrast, other observers (Lincoln, 1990; Vermillion & Pfeiffer, 1993) predict that staff members themselves will more frequently use evaluation methods to develop improvements in programs. When evaluation systems are developed in ways which clearly distinguish between information collected for (1) the purpose of program improvement and (2) the purpose of individual performance appraisals for salary increases, conflicts among evaluation, service delivery, and administration may be minimized. That happy day has not yet come, but when evaluation is built into program management and all agree that program improvement is both possible and desirable, the potential for conflict will be reduced (Donaldson, Gooler, & Scriven, 2002).

Activities Often Confused with Program Evaluation

Sometimes it is easier to understand a concept when one understands what the concept does not include. Program evaluation is often confused with basic research, individual assessment, and compliance audits. Although these activities are valuable, when program evaluation work is confused with one of these other activities, the evaluation becomes all the more difficult to carry out.

Basic research concerns questions of theoretical interest, without regard to the information needs of people or organizations. By contrast, program evaluators gather information to help people improve their effectiveness, to assist administrators to make program-level decisions, and to enable interested parties to examine program effectiveness. Evaluators, of course, are very interested in theories of why a service might help its participants. Understanding theories helps in planning programs and selecting variables to observe. However, contributing to the development of theories can only be a delightful side benefit of a program evaluation. Evaluation findings should be relevant to the immediate or short-term needs of managers and must be timely. If program staff members believe that evaluators are collecting information primarily to serve research interests, cooperation may well be lost.

Human service staff members often confuse program evaluation with the assessments made by educational psychologists, personnel workers, and counseling psychologists who administer intelligence, aptitude, interest, achievement, and personality tests for the purpose of evaluating a person's need for service or measuring qualifications for a job or a promotion. These activities are not part of the work of a program evaluator. In program evaluation, information about the level of job performance, educational achievement, or health may well be gathered; however, the purpose is not to diagnose people, determine eligibility for benefits, or choose whom to hire or promote. Instead, the purpose is to learn how well a program is helping people improve on those variables.

Last, the methods and objectives of program evaluation differ from those used by program auditors examining government-sponsored programs to verify that they operate in compliance with laws and

regulations. When Congress supports a program, it is important that the funds be spent as Congress intends. If the appropriation was for the enrichment of elementary schools, spending the funds for high school laboratories constitutes fraud. If 10,000 students were to be served, documentation of services for close to that many children ought to be available. Program evaluators are concerned that the program serve the right number of children, but in addition, evaluators would be particularly interested in how the services have affected the children.

The work of program auditors is closely related to accounting, whereas program evaluators tend to identify with education and other social sciences. Auditing and evaluation have moved toward each other in recent years because neither can give a complete picture of a program (Hatry, Wholey, & Newcomer, 2004). The change of the name of the GAO from the "General Accounting Office" to the "Government Accountability Office" illustrates this change. Nevertheless, if a program evaluator seeking to help a program make improvements in its services is viewed as a program auditor seeking to verify compliance with laws, it is easy to imagine that the evaluator's purpose could be misunderstood.

Evaluation and Related Activities of Organizations

Some agencies may combine evaluation with another activity of the agency. Five functions that sometimes include evaluation are research, education and staff development, auditing, planning, and human resources.

Research

Some human service organizations routinely sponsor research directly with operating funds or indirectly through grants. For example, a police department was awarded a grant to study the effect of providing emotional support to victims of crimes. The social scientists responsible for such research would be excellent colleagues for an evaluator. University-affiliated hospitals carry on research; researchers working with psychiatric problems or emotional issues related to medical care deal with research problems sufficiently similar to those faced by evaluators to serve as colleagues.

Education and Staff Development

Some organizations have joined the functions of education and evaluation. There is a long-standing precedent for this marriage in educational psychology. New study materials and curricula are evaluated before being widely used by school districts. Universities have combined faculty development and program evaluation into one office. Businesses have developed major educational programs to train employees to qualify for more responsible positions; such programs need monitoring.

Auditing

As mentioned earlier in this chapter, some states have combined the functions of program evaluation and program oversight often carried out by state auditors. There are important differences in the assumptions and roles of auditors and evaluators, but there is precedent for coordinating these efforts in the inspector general offices in the federal and state governments.

Planning

Planning is an activity in which people with skills in program evaluation are needed because many service providers are not data oriented, feel uncomfortable constructing surveys, and may be unable to analyze the information obtained. Evaluators can contribute to the task of estimating the level of community acceptance. Sometimes evaluators can help planners simply by posing probing questions about the assumptions implicit in the plan or about the steps involved in its implementation.

Human Resources

Large businesses have offices of human resources. The office oversees all hiring, compensation, and professional development of employees. Policy development, compensation plans, training programs, and management effectiveness seminars all require evaluation if an organization is to be as effective as possible.

CASE STUDY 1

Open Arms

In May 2014, REACH Evaluation presented its program evaluation report of the Open Arms Children's Health Dental Service, a program at Home of the Innocents. The report is a good example of a comprehensive evaluation. Open Arms involved on-site dental services for special needs children supported with funds from the Social Innovation Fund through the Foundation for a Healthy Kentucky and in partnership with the University of Louisville School of Dentistry's Pediatric Dental Program.

Meeting Needs

The extensive three-page section "Documentation of Need" included a summary of research that "linked poor oral health to illness, chronic disease and early mortality" (p. 9), research on the wide span of health needs among foster children, and the comparative point that Kentucky "has the highest percentage of edentate persons (those who have lost all their natural teeth due to tooth decay or gum disease) ages 18 to 64" (p. 9) along with multiple additional points of evidence regarding the need that Open Arms was intended to address.

Implementation

The issue of what was actually done in the program was addressed at several points in the report. For example, copies of medical notes for dental visits for the same child before and after the program was initiated show enormous differences in detail—about two pages of very specific notes after compared to one line on a card before. In the "Early Lessons Learned" section, the progression of staffing strategies during the program was described, detailing how initial plans were adapted.

Stakeholders

The evaluators had contact with clients, client family members, and permanent staff as well as dental students and dental residents who received training as a part of the program. As with many evaluation reports, the category of "stakeholders" was not listed explicitly, but the perspectives of multiple stakeholders were clearly considered as a part of the evaluation. For example, although programs generally focus on clients, and clients' perspectives were included, so were accounts of some parents of clients, who are also important stakeholders.

Side Effects

As with stakeholders, the report did not explicitly note that it found side effects. But with the stated goal of improving children's dental care, the report's note that having dental services provided on site saved staff time in several ways was an implicit description of an unanticipated side effect. Specifically, reduced time was needed to transport children to dental services because the services were on site, and they found greater efficiency when scheduling visits because the person planning the times for the dental services had access to information about children's other activities, and was thus able to avoid potential conflicts more easily. No negative side effects were noted.

Improvement Focus

Suggestions by staff and clients for ways to improve the program were requested and listed. In addition, the evaluators showed an implicit focus on improvement by suggesting strategies for addressing one of the central challenges for the program—how to continue it past the initial period of outside funding.

Outcomes

As the primary goals of the program were fairly straightforward—serving 880 children a year, serving special needs children, and serving 80 dental students and 6 pediatric residents, the report appropriately presented direct counts of the relevant groups as the primary outcomes. For two 12-month periods (overlapping, because the available data covered only 21 months) 920 and 1,411 children were served. The 1,475 special needs individuals served during the 21-month period were described with appropriate categories, such as "Medically Fragile" or "Foster Care." And the planned number of students and residents was met. The results of inferential statistical analyses showed a significant improvement in treatment acceptance over time, while also noting that the changes were relatively modest, partly because the scores were very high to begin with (a ceiling effect). They also found very much greater improvement in four groups of children (such as Medically Fragile and Refugee) than in four other groups (such as Outpatient/Nonmedically Fragile). Several summaries of qualitative data addressed issues not analyzed quantitatively, such as the improvement in the medical notes indicated above.

Nuances/Mechanisms

Although no formal statistical analyses were conducted to determine mediators or active ingredients, it was clearly noted that the attitudes and behavior of the staff, such as being "very relaxed" and being very much more patient than was expected, played a key role in the success of the program. In addition, the "variety of behavioral aids and techniques" (p. 4), designed to help with the challenge of providing dental services for special needs children, appeared to be another important component.

Summary

As noted at the beginning, REACH's evaluation is a good example of a comprehensive program evaluation report. The report itself is a high-quality, 50-page booklet with pictures, graphs, and appendices for the tables reporting statistical analyses typically skimmed at best by most program stakeholders. As you will see through the rest of this text, there are many quite good program evaluations that do not address all aspects of MISSION as thoroughly or as well as this evaluation, yet still provide substantial information about the program to those who want to know as well as important feedback for those who wish either to improve the original program or adapt it to other settings.

Sources: REACH Evaluation, *Open arms,* Louisville, Author, 2014; B. Birkby (personal communication, September 13, 2017).

EVALUATOR PROFILE 1

Ben Birkby, PsyD—Not Part of the Original Plan

When Ben Birkby was applying to doctoral programs in Clinical Psychology nearly 25 years ago, he was focused on preparing for clinical work with children and families. He recalls now that he knew very little about program evaluation, let alone imagined that it might become an important part of his career. Events would turn out differently than he expected, however. He enrolled in the PsyD program at Spalding University in Louisville, Kentucky. Having indicated that he wanted to be considered for financial assistance, his record of involvement in undergraduate psychological research caught the notice of Dr. Bob Illback, a professor at Spalding, who had developed a Program Evaluation specialty at REACH of Louisville, a local agency also providing clinical services to children and adolescents.

Ben started work as an assistant to Bob for program evaluations on his first day of graduate school, and looking back, he notes that learning program evaluation was interwoven into his entire graduate program. He quickly realized that, unlike his peers, he was gaining valuable experience in this less-well-known part of the field, finding a crucial complement to his clinical work, and building an important foundation for a component in his career. Following graduation, Ben went to work at REACH of Louisville where, for nearly 20 years, he has worked clinically with children, adolescents, and families while also playing a major role in REACH Evaluation—an organization integrated with REACH of Louisville that is dedicated to improving services and bringing about organizational and community change through program evaluation. Over time, he has had a hand in nearly 100 evaluations of varying sizes and scopes and now helps lead a team of 11 staff—a long way from knowing almost nothing about the area.

Two important aspects of Ben's approach to clinical work are developmental and systems perspectives—paying attention to expected patterns of change over time and the recognition of multiple levels of perspective. Especially with adolescents, attending to and intervening with not only the individual but also the family, the school, and more, is essential. Ben said, "Program evaluation is critical to my clinical work—it allows me to live out what I say is my professional identify as a systems clinical psychologist." In addition, as he likes doing many different things and learning about different things, program evaluation is a good fit for those interests. As he has to learn and do a wide range of things in evaluations, both as a clinical psychologist and personally, he feels he is "living his dream." Finally, he noted that he sees evaluators serving a crucial role; they are "the people in the background working to insist on reason and quality data in our programs."

Source: B. Birkby (personal communication, September 18, 2017).

SUMMARY AND PREVIEW

This chapter outlines the major questions and purposes of program evaluation: assess unmet need, document implementation, consider as many stakeholders as possible, look for positive as well as negative side effects, consider how to improve the program when possible, measure outcomes, and examine what can be learned about why the outcomes occur. Such activities are undertaken to help plan and refine programs, to assess their worth, and to make corrections in ongoing operations. Some ways in which evaluators fit into the activities of agencies are described.

Chapter 2 deals with steps in planning an evaluation. Working with program staff and sponsors will often reveal that some people fear program evaluations; when an evaluation must have the cooperation of staff, it is necessary to allay these fears before beginning.

STUDY QUESTIONS

1 One of the interesting aspects of program evaluation is its relevance to a variety of activities and agencies. Gather some examples of activities related to program evaluation from newspapers or news magazines. You might find material on medical care, education, or public policy in any area. The material might concern advocacy of new proposals or criticisms of current activities. Consider what information might illuminate the debates.
2 Some people have suggested that while people want to believe that something is being done about social problems, most people don't really care whether programs have measurable impacts on the problems. In other words, social programs are symbols, not real efforts. Consider the validity of such comments in the context of the reports you found in response to the first question.
3 Illustrate how program evaluation activities could be applied in an organization with which you are familiar.
4 List the advantages and disadvantages of making program evaluation a part of human service delivery systems. Save your lists; when you are finished with the course, consider how you might change them.
5 Compare your ideas on evaluating the Sexual Assault Prevention Program mentioned in the introductory paragraph of this chapter with those of other students. What reasons lie behind different choices of outcomes considered relevant?

ADDITIONAL RESOURCE

W. K. Kellogg Foundation. (2010). *Evaluation handbook*. Retrieved from https://www.wkkf.org/resource-directory/resource/2010/w-k-kellogg-foundation-evaluation-handbook.

This work provides a helpful overview and succinct details of the program evaluation process. In addition, it highlights some related points such as the way that values and assumptions affect evaluation work. Best of all, like many resources on the web, it is free.

2 Planning an Evaluation

OVERVIEW

Preparation precedes production; it is often harder to plan well than to follow a good plan. Experienced evaluators know that the time put into planning an evaluation is well spent. Program evaluations begin in a variety of ways. Formal Requests for Proposals are sometimes published to invite a plan for an evaluation. Alternatively, program personnel may initiate an evaluation, the central administration or a funding agency may seek an evaluation, or public dissatisfaction might prompt government officials to call for an evaluation.

Regardless of who initiates an evaluation, there are a number of important tasks to complete. Evaluators need to become familiar with the nature of the program, the people served, and the goals and structure of the program, and, above all, learn why an evaluation is being considered. Evaluators seek to meet with program personnel, program sponsors, and other stakeholders, some of whom may question the need for an evaluation. Then, in consultation with these groups, evaluators may suggest how best to evaluate the program. These consultations may well involve negotiating the timing of an evaluation, the manner of conducting it, and its costs. Points such as whether the evaluator is internal or external to the organization will affect some of these processes such as how one learns about stakeholder perspectives and building relationships with stakeholders. Additional elements such as the population and problem(s) addressed by the program, the goals for a given evaluation, and other characteristics also affect details of preparing for an evaluation. Further tasks include deciding on the elements of the program to be assessed and on what data to use. Perhaps sufficient data have already been collected, or perhaps new data will need to be gathered as part of the evaluation. The analysis strategy needs to be determined as well as how and when and to whom reports of the evaluation will be presented.

It is important to understand that these tasks cannot generally be completed in a simple linear fashion—the results of some of the elements listed later may require returning to earlier elements. For example, talking initially with staff, an evaluator might plan to use data from client records to determine outcomes. But examining the records may show problems with that strategy, such as there being too many cases of missing data or that the information is not in a form to answer the evaluation questions. The evaluator would then meet with staff again to find a better solution.

One aspect of planning an evaluation, the choice of the overall approach to the evaluation, called the model (or sometimes a combination of models), generally occurs after some of the initial steps such as learning about the program and meeting with stakeholders. But models are conceptual guides, and the decision of which one or ones to use is informed by a number of other elements in the evaluation process. A description of models and a summary of the main options is presented first to clarify these conceptual points. Then the basic steps of preparing for an evaluation are described, most of which will be elaborated in greater detail in the later chapters. This chapter's case study, in addition to providing another real-life example of an evaluation, also illustrates how models are choices made by evaluators that affect important aspects of evaluations. In the summary at the end of the chapter, a simple checklist of the main elements when preparing for an evaluation is provided as a guide, especially for new evaluators. As might be expected, there are many relevant resources available online to help with various aspects of planning an evaluation. See the eResources for an annotated list with active links to websites helpful for planning program evaluations. Looking further ahead, Chapter 3 addresses program theory and the selection of criteria for the evaluation, while Chapter 4 provides a detailed discussion of choosing specific measures.

EVALUATION MODELS

There are many different approaches to program evaluation; Stufflebeam and Coryn (2014) list 23 different models, and at least 11 more have been suggested (Patton, 2016; Posavac & Carey, 2007; Pouw et al., 2017; Sturges & Howley, 2017; Thomas & Parsons, 2017). Some are used more commonly than others, but all approaches have something to recommend them, otherwise evaluators would not have used them. Although some of these models of evaluation seem to differ greatly, in fact, the different approaches are all similar in being ways to shed light on the value and worth of programs through a range of methods or from different perspectives. The models differ primarily in the initial approach of the evaluator in gathering evaluation-relevant information. As there are far too many models to cover adequately in this printed text, the essential features of models are described, with explanations of specific models that illustrate four general approaches and a table listing some features of those models. An extensive list with descriptions and a correspondingly much larger table summarizing important features of those models is available in the Chapter 2 folder of the eResources.

Much like different theoretical orientations or research strategies, the models or approaches reflect different assumptions about important aspects of evaluations. Along the lines of the relatively simple natural evaluations described in the first chapter, one rather simple model is the BLACK BOX. Those following this model make few assumptions and take an almost stereotyped "bottom line" approach, concentrating on outcomes without reference to much else. Such an evaluation considers whether program participants are better off following the program than they were before without examining anything else about the process such as why they are better off or why some participants improve more than others. To be fair, some Black Box evaluations at least specify how much better participants need to be in order to count as successes, and some evaluators might say they are using a Black Box model, but still address other matters such as problems with implementation if they are substantial enough. But the simplicity of this model means that it is most commonly seen as a minimum standard to be exceeded (Mead, 2016; Peck, 2016; Solmeyer & Constance, 2015). In general, this text presents a sustained rationale not to settle for a Black Box model.

In contrast, in many educational settings, especially accreditation of a program, evaluators must start by learning what that particular institution intends to achieve by looking at the specific objectives noted in formal documents—using an OBJECTIVES-BASED model. The goal of the evaluation is then to determine the degree to which the program achieved those objectives. In those settings, it is inappropriate for evaluators to criticize established objectives—they are taken as givens. For example, although many colleges and universities include philosophy as an important element of liberal arts studies, some colleges do not consider it to be an appropriate part of their educational plan. An accreditation visit using an Objectives-Based evaluation model would not describe the lack of philosophy classes as a problem at such a school, whereas an evaluation of a college that intended to teach philosophy but failed to do so would have the deficiency noted. To be clear, attending to organizations' objectives is not only seen in Objectives-Based evaluations—all good evaluators should be aware of the stated objectives of the program. But whereas other models indicate additional bases for determining appropriate outcomes, Objectives-Based evaluations start and stay with the organization's own goals.

A THEORY-BASED model explicitly includes elements that broader research has identified as important components of programs like the one being evaluated. At a basic level, this means that the selection of outcome measures is informed by current theoretical understandings of the relevant field, not just the specific objectives of the organization. For example, a basic understanding of Major Depressive Disorder includes impairment of subjective affect (feeling depressed) and the presence of physical symptoms such as greatly increased sleep and change in appetite. An agency that requested a Theory-Based evaluation for a program treating depression but wanted evaluators only to measure clients' subjective feelings, and not physical symptoms, would face a conflict.

Theory-Based models typically also consider other elements indicated by current research, including portions of the interventions that are believed to be some of the mechanisms for change. Recall the example at the beginning of Chapter 1 regarding the counseling center's program to reduce sexual assault. One focus of expected mechanisms might be training for women and men in behaviors intended to reduce their risk of being sexually assaulted as well as measures of how frequently trained students show those behaviors. Another focus based in current theory would look at training for bystanders to intervene in situations with the potential

for sexual assault, such as when some partygoers show signs of decreased functioning as a result of drinking alcohol (Hahn, Morris, & Jacobs, 2017; Koelsch, Brown, & Boisen, 2016; Senn & Forrest, 2016). The explicit point of attending to theoretical insights in this model is also a common contribution that evaluators can make to programs—many programs continue to use somewhat dated approaches, and good evaluators can help program staff learn about newer insights.

A model emphasizing a different perspective is the EMPOWERMENT model, sometimes called a CAPACITY-BUILDING approach. Whereas the above models address assumptions about the appropriate outcomes and mechanisms to be measured, an Empowerment evaluation stresses important aspects about the process of the evaluation, especially the interactions among evaluators and various stakeholders. Another way to summarize the Empowerment or Capacity-Building approach is that instead of primarily providing the product of an evaluation of a program, these evaluators provide the service of training agency staff how to evaluate their own program while working on the product of the evaluation. In contrast to an Objectives-Based approach that would roughly adopt agency goals and outcomes, and also in contrast to a Theory-Based approach that would roughly draw appropriate goals and outcomes from research in the appropriate field, Empowerment approaches emphasize collaboration—evaluators and stakeholders work together to determine the goals and outcomes of the evaluation.

Figure 2.1 notes some of the basic points of these four models, which provide a window into a much larger set of models. For example, other models focus on the method of gathering information. One common technique is to use established assessments such as symptom checklists for treatment of mental disorders or knowledge as revealed on standardized educational tests. But other settings and evaluators consider that the stories clients tell will convey rich information about many relevant issues that cannot be reduced to numbers along a single dimension such as distress. In those cases, evaluators typically collect qualitative data. Several models are based on different emphases regarding data collection.

Model	Brief Description	Strengths	Weaknesses
Black Box	The focus is exclusively on outcomes, not considering mechanisms or other explanations why the outcomes occur.	The simplicity is attractive to some stakeholders. Evaluations can be completed more quickly and less expensively.	This approach is increasingly seen as too simplistic and inadequate to provide necessary insight into important elements such as mechanisms.
Objectives Based	This approach accepts the objectives of the organization as the outcomes without critique. It is especially common in accreditation of educational institutions.	It is seen as particularly fair because programs are evaluated on the basis of what they explicitly claimed they intended to accomplish.	In many services, there is general agreement of what is valuable regardless of whether one organization agrees. This approach does not address agreement on goals in the field.
Theory Based	Appropriate theories from the relevant literature inform this evaluation—indicating the expected outcomes and mechanisms.	There is a good match to current research and evidence, which raises expectations for excellence.	To the degree that the organization was unaware of the most recent research or unable to change to address it, stakeholders can feel that the standards were switched on them.
Capacity Building / Empowerment (These two labels are sometimes considered separately, but they are often seen as related if not the same.)	One primary purpose is to conduct an evaluation in collaboration with agency staff to demonstrate Program Evaluation knowledge and skills with a goal of helping staff be able to conduct some level of future evaluations.	Collaboration often yields more comprehensive information and insights. Training internal evaluators strengthens future agency work.	Depending on how empowerment is handled, agency evaluators could over-emphasize positive findings and downplay or even ignore negative ones.

Figure 2.1 Characteristics of some basic evaluation models.

There are a couple of general points to make here. First, although a few models are incompatible with certain other models, most provide an emphasis that can be combined with one or more other approaches. To use the analogy of tools, many tasks require using a set of tools rather than just one. As valuable as a hammer may be for one part of a job, a screwdriver may also be needed for another part, and neither tool can substitute well for the other. Skilled evaluators are able to adjust to different settings or demands by employing a different model or combination of models, even if they have their typical approach, just as people repairing things adjust to the specific needs by picking the appropriate tool or tools. Second, the choice of the model or models depends on a number of factors. The particular agency and program being evaluated, the reason for the evaluation, the general skill set and approach of the particular evaluator, and other elements such as time frame and resources available may all affect the choice of the model. Hopefully it is clear that because so many elements are relevant, consideration of the desired model needs to be a part of the very early stages of any evaluation.

An Improvement-Focused Approach

Regardless of the other models chosen, however, this text presents an approach to evaluation in which

program improvement is the overarching attitude. Improvements can be made in programs when discrepancies are noted between what is observed and what was planned, projected, or needed. Evaluators help program staff to discover discrepancies between program objectives and the unmet needs of the target population, between program implementation and program plans, between expectations of the target population and the services actually delivered, or between outcomes achieved and outcomes projected. Finding discrepancies is not a perverse trait of evaluators. If the point of evaluation is to improve the program—which is nearly always the point—then discrepancies provide a place to seek to effect improvements. To learn how to improve a program, staff members need to discover what has occurred as planned and what has not. Do the clients have unmet needs that were anticipated or different ones? Do the staff members have the needed skills or has it been difficult to teach them the needed skills? Has the promised support materialized? Does the operation of the program suggest that the conceptual basis of the program is sound or that it can be improved? To deal with such questions, objective information is needed, but qualitative information can often help with interpretation. Personal observations provide direction in selecting what to measure and in forming an integrated understanding of the program and its effects.

The improvement-focused approach, it is argued, best meets the criteria necessary for effective evaluation: serving the needs of stakeholders, providing valid information, and offering an alternative point of view to those doing the really hard work of serving program participants. To carry this off without threatening the staff is one of the most challenging aspects of program evaluation. Further, this approach is not to be interpreted as implying that evaluators are not to celebrate successes. After all, evaluators are to look for merit and worth in programs. Managers of programs are going to pay close attention to evaluation reports when evaluators demonstrate that they recognize the strengths of programs, not just the ways in which the program has fallen short and can be improved.

As noted above, sufficient description of even most of the major models requires more space than available in this textbook. See the eResources for extended explanations and a summary table of models, their main characteristics, and the most obvious strengths and weaknesses. Further, new models introduced to the field since this edition was published will be added on at least a yearly basis.

The practical steps of planning a program evaluation are described next. Some of the steps will be different depending on specifics such as whether it is an internal or an external evaluation, while other steps will be similar across most settings.

Steps in Preparing to Conduct an Evaluation

In considering how to begin preparing for an evaluation, Sarason's (1972) points about "before the beginning" apply both to the evaluation itself and to the program being evaluated. Interested readers will find Sarason's chapter by that name still highly relevant. In brief, he cautioned against ignoring the context of any enterprise. Programs were created in specific settings because of particular forces and people. And although many of the motivations to create programs may have been based on increasing services for those in need, at least sometimes programs are created in reaction or opposition to other programs. Likewise, evaluations are not requested only because all stakeholders want to gain comprehensive insights to make services better. The point here is not to create suspicious evaluators looking for ulterior motives, but as a reminder that there may be important history to consider with regard to programs and evaluations.

With that point in mind, while planning an evaluation in an applied setting focused on emotionally sensitive issues—such as health, justice, and work success—it is necessary that the communication be clear and agreements made be explicit. Oral agreements between evaluators and administrators should be followed up with memos outlining the discussions and the evaluator's understanding of the agreements. Depending on memory is very unwise; evaluators need to recognize that since program administrators deal with many problems during each day, the evaluation merits the administrator's attention during only brief moments during a week. Putting agreements on paper serves to remind all parties about the plans and is a record if disagreements occur later. Furthermore, reviewing developing plans described on paper can suggest implications that neither evaluators nor administrators had considered previously. A common reality late in the process of an evaluation is regret that some details were not specified in writing earlier.

Identify the Program and Its Stakeholders

Obtain a Complete Program Description

The first thing effective evaluators do is to obtain descriptions of the program. It makes quite a difference whether an evaluation is proposed for a new program or a well-established one; whether it is locally managed or offered at many sites; whether people seek to participate voluntarily or are assigned to complete it; whether participants are functional or suffer cognitive, emotional, or physical problems; whether it serves 25 or 25,000 people; and whether the program theory is well-articulated or based on common sense.

Meet with Stakeholders

The second thing effective evaluators do is to identify the stakeholders. Stakeholders are those people who are personally involved with the program, derive income from it, sponsor it, or are clients or potential recipients of the program's services (Bryk, 1983; Sieber, 1998). Depending on the setting, people who live near program locations may be stakeholders.

Program personnel are usually more personally involved in the program than either the sponsors or the clients. The program director is a key person with whom the evaluators relate during the entire project. It helps to learn as much as possible about the director's training, program philosophy, vision for the program, and reputation. The people who deliver the services need to be involved as well. Such involvement should begin early so that they can add insights that only they can provide because only they know what happens hour by hour. A consensus between the director and the staff about the purposes and methodology is also critical. Such cooperation increases the chances that all groups will cooperate in planning the evaluation, collecting data, and finally using the findings.

Program sponsors should be considered. At times program personnel are the sponsors; in other situations sponsors are foundations, government agencies, or central administrators of the institution housing the program. Often specific individuals are responsible for the evaluation. For example, a member of the school board or the vice president of a hospital might be the person who must be satisfied with the evaluation. It is important that evaluators meet with sponsors early during the planning phase to answer questions about the evaluation; however, such people will seldom be involved in detailed planning. As the evaluation proceeds, sponsors appreciate progress reports. Suggestions for keeping sponsors informed and helping sponsors use the information gained through the evaluation are given in Chapters 13 and 14.

The clients or program participants also need to be identified. The type of contact with clients will vary with different types of evaluations and programs. Evaluations of services directed to whole communities might require contact with just a small sample of eligible residents, those of school-based programs might depend on work with parents, and evaluations of small programs might involve all participants in some way. Having a good understanding of the needs of participants is necessary because, after all, it is for their welfare that the program is being provided. Although these categories cover most common groups of stakeholders, there may be others, depending on the specific program.

Become Familiar with Information Needs

After learning about the program and meeting stakeholders, evaluators should ask: (1) Who wants an evaluation? (2) What should be the focus of the evaluation? (3) Why is an evaluation wanted? (4) When is an evaluation wanted? (5) What resources are available to support an evaluation?

Who Wants an Evaluation?

Ideally, both program sponsors and program personnel want the program evaluated. In such situations evaluators usually interact with cooperative people secure in their professional roles who want to verify that the program meets their expectations and who wish to improve or extend the program.

When the sponsors initiate an evaluation without explicit agreement by personnel, evaluators are faced with helping the program personnel become comfortable with the goals and methodology of the evaluation

before data collection begins. If this effort fails, evaluators face the possibility of active opposition or passive resistance; either way, they will not gain the cooperation essential to carry out the project (Donaldson, Gooler, & Scriven, 2002). When program personnel treat an evaluation as a means of improving the effectiveness of their work, they are more likely to give the evaluators assistance in data collection and to volunteer valuable insights into the interpretation of the data.

Alternatively, if the program personnel initiate an evaluation, evaluators need to be sure that sponsors are convinced of its usefulness. Sponsors who are disinterested in the evaluation during the planning stages are not likely to pay attention to the findings or to support improvements that might be recommended.

What Should Be the Focus of the Evaluation?

During meetings with the sponsors and personnel, evaluators often learn that the term *program evaluation* does not have the same meaning for everyone. Although it would be better for them to think in terms of formative evaluations that would help them to retain positive features of the program and modify or improve other aspects of their work, sometimes program personnel merely seek affirmation of their current efforts. Program sponsors may want a summative evaluation if they are under pressure to divert resources to another program. Others might confuse program evaluation data collection instruments with measurements of individual accomplishments similar to employee appraisals (Bauer & Toms, 1989).

At this point, evaluators help stakeholders to learn what types of evaluations best meet their needs and resources. The choice is seldom between one type of evaluation or another, but rather a matter of emphasis of two or more types. Incorporating some elements of various types of evaluations is often desirable, depending on the complexity of the goals of the program and the resources available for the project. A number of issues can be addressed in the overall evaluation such as the degree to which the target audience is reached, a program achieves the planned objectives, the achieved outcomes match the unmet needs that led to the program, and the benefits exceed costs. In addition, the specific aspects of the program that fostered achievement of the outcomes and the value of program outcomes can be examined as well.

Why Is an Evaluation Wanted?

Closely tied to the previous question is the issue of why evaluation is wanted. People seldom commission an evaluation without a specific reason. Effective evaluators put a high priority on identifying the reasons why an evaluation is wanted. Are there some groups in the organization who question the need for the program? Is the foundation that provided the funds expecting an evaluation? What level of commitment is there to use the results of an evaluation to improve decision making? Ideally, program personnel seek answers to pressing questions about the program's future. How can it be improved? Are we serving the right people? Should we expand? Evaluators expect that different stakeholder groups have differing priorities. Some want a smoothly functioning program, others want tips on making their work more effective, and still others want to see services expanded. One aspect of an evaluator's role is to help both the sponsors and program personnel arrive at a consensus about the purposes of an evaluation.

Some reasons for evaluating programs are undesirable. For example, an administrator may use program evaluation as a ploy to postpone or avoid making a decision. Evaluations are also inappropriate when administrators know what decision they will make but commission a program evaluation solely to give their decision legitimacy. Effective evaluators attempt to help administrators develop better reasons, when possible.

When Is an Evaluation Wanted?

Stakeholders often want an evaluation completed quickly. Part of the planning process involves agreeing on a balance between the preferences of the stakeholders and the time needed for carrying out the best evaluation possible. When setting a viable completion date, evaluators must consider the time it takes to prepare a proposal, access records, arrange meetings, develop measures of the criteria for intermediate and long-term outcomes, find and observe a reasonable number of program participants and sites, and draw together the observations and prepare written and oral reports. Internal evaluators are often restricted in how many

observations they can make because schools, hospitals, and businesses include only a limited number of students of a certain age, patients of a given diagnosis, or employees in a specific job. Furthermore, the type of evaluation requested leads to different constraints on the time needed to plan and complete an evaluation.

There is no formula to determine how long a project will take. With experience, however, evaluators develop a sense of the relationship between the scope of a project and the time needed for its completion. A good technique is to break the overall project down into small steps and then estimate how much time each step in the project needs. Using the list of considerations in the previous paragraph might help in estimating the time needed to complete an evaluation. On the other hand, the selection of a completion date might be specified in a request for proposals; in that case, evaluators adjust their proposed timetables to this date.

What Resources Are Available?

Besides time, another factor that can limit an evaluation is the availability of resources. Grants typically include a specified amount for an evaluation. Of course, internal evaluators cannot be reckless with resources either. The assistance of program personnel can hold down the expense of data collection. Even if no formal contract is to be signed, listing in writing what is to be done as part of the evaluation is advisable.

Assess the Evaluability of the Program

After coming to understand the program and learning what information the stakeholders need, evaluators must consider the resources available for meeting those needs. This process, called an evaluability assessment (Wholey, 2004), is intended to produce a reasoned basis for proceeding with an evaluation. It is best if the stakeholders agree on both the objectives of the program and the criteria that would reveal when successful outcomes have been achieved. If stakeholders cannot agree on what makes the program a success, evaluators are faced with a particularly difficult situation because they would need to provide a greater variety of outcome data and may be unable to draw clear conclusions.

A program is also not ready for a thorough evaluation until its conceptual basis has been developed. One review showed that less than 30% of published evaluations described programs with theoretical formulations linking the program to the desired outcomes (Lipsey, Crosse, Dunkle, Pollard, & Stobart, 1985). Others agree that many programs do not have an explicit statement of how the program is expected to affect people (see Bickman, 2000; Cook, 2000; Leeuw, 2003). Upon questioning, it may turn out that some implicit program theories are no more complicated than "we tell them what to do, and they do it" (see reviews by Posavac, 1995; Posavac, Sinacore, Brotherton, Helford, & Turpin, 1985). But knowledge about the health effects of regular exercise and a balanced diet low in fat, refined sugar, and salt is not sufficient to motivate most people. Many steps intervene between the provision of knowledge and actions; such steps can include social support, reinforcing appropriate behaviors, reminders, skill in applying knowledge, belief in the personal applicability of information, and so forth. Discussions with evaluators can reveal that planners had not developed a clear rationale on which to base a program or do not have the freedom to make changes in response to an evaluation. Recognizing such situations even before gathering any data is a contribution that evaluators can make.

When deciding whether a program has a useful theoretical basis, it is helpful to construct an impact model showing how the elements of the program lead to the expected changes in the program participants (Lipsey, 1993; Renger & Titcomb, 2002). Impact models are mentioned at several points in this text. In brief, an impact model is a diagram listing the activities making up a program on one side of a page and the outcomes the program is expected to produce on the other, with additional elements depending on the style of impact model. See Figure 2.2 for an impact model of a hypothetical program to improve the lives of children in high-risk neighborhoods. The program is based on the idea that the children are not learning good conflict resolution skills and show limited respect for authority. Different assumptions would be reflected in a different impact model. The program components or activities are at the top. What happens during the program sessions, outputs, is listed next. The intermediate outcomes are given first; it is believed that they must occur in order for the ultimate outcomes to be achieved. Finally, at the bottom are the long-term, desired impacts.

Impact models, such as the one shown in Figure 2.2, help the evaluator to decide what to observe and

measure. The best evaluations involve assessing the quality of the program's activities (i.e., faithful and high-quality implementation), the degree to which the intermediate outcomes were achieved, and the extent to which the students achieved the outcomes at the end of high school. If one focused only on the objectives for students (the points in the lower section) and did not find improvements compared to the children who did not participate in the social learning program, one would not know what went wrong. Figure 2.2 is a very simple impact model, but it still is valuable because it prompts the stakeholders to look at the elements of the program and trace the process by which the outcomes are expected to be obtained. It also tells the evaluator what to consider measuring. (See the eResources for other examples and the W. K. Kellogg Foundation Logic Model Development Guide, 2004 for a very helpful resource.)

Plan the Evaluation

Once it is decided to conduct an evaluation, planning begins in earnest. Examining published research and evaluations of similar programs, determining the best methodology, and preparing a written proposal complete the preparation phase.

Examine the Literature

When evaluators work in an area that is new to them, making a careful search of the literature before designing or developing new instruments is important. Evaluators can learn from the successes and failures of others and get a picture of the methodological, political, and practical difficulties that must be overcome. Searching published materials is much easier than in the past. The research literature can be accessed through Internet searches in libraries and even from home. For medical, mental health, and educational program evaluations, MEDLINE, PsycINFO, and ERIC are readily available. Many other specialized systems have also been developed that have taken much of the drudgery out of literature searching. Once some reports are found, the bibliographies of these articles provide additional references.

Impact Model for a Social Learning After-School Program for Middle School Children in High-Risk Neighborhoods

Statement of Objectives: Increase school achievement and aspirations, immunize students against negative community influences, develop greater mutual respect among peers, and develop increased acceptance of adult guidance.

Activities	**Staff**
Gym time with game leaders **Study time** with leaders and volunteers **Snack** (healthy foods) shared with leaders and volunteers **Life skills training** with role playing (planned by Youth Leader, modeled by high school students and adult volunteers)	**Youth Leader** (master's degree in guidance) Successful, **high school students** from the neighborhood **Adult volunteers** from the neighborhood

Outputs	Healthy social climate modeled during gym time and training	Tutoring in Math and English provided during study time, other tutoring as needed		
Intermediate outcomes	Attends school regularly (less truancy)	Pays closer attention in classes (less classroom misbehavior)	Does better school work (higher grades)	Gives less approval of gang activity (attitude change)
Outcomes	Participates in follow-up program in high school	Develops healthy aspirations for education and life	Graduates from high school	Does not associate with gangs during high school
Impacts after graduation	Obtains job	Attends trade school, community college, or college	Avoids delinquency, avoids crime	

Figure 2.2 Simple impact model for an after-school program for middle school children.

While reading reports of previous work, evaluators keep several key questions in mind. In what ways is the program evaluated similar to the program whose evaluation is being planned? What research designs were used? Can some of the measures of the outcome criteria be adapted? How reliable and valid were the measures? What statistical analyses were used? Is there a consensus among the reports? If there are conflicting findings, are these conflicts due to different community characteristics, compromises in selecting samples, variations in research design, or variations in how the programs were administered? What issues were not addressed?

Plan the Methodology

After reviewing the literature, the evaluators are ready to make decisions regarding sampling procedures, research design, data collection, and statistical analysis. Later chapters are devoted to some of these topics, but it is helpful to preview the main issues now to show how these decisions contribute to planning an evaluation.

Once the program's target population has been identified, evaluators need to consider whether to use the entire population, a random sample, or a sample selected for other reasons. One argument for including the entire population is political: some people may be offended if they are not included. Evaluations within a large organization will be more readily accepted if all employees have an opportunity to participate rather than only a random sample. If a state auditor's office is evaluating a statewide program, gathering information from

every county, or at least from every state representative's district, would be wise.

When there are concerns about particular types of participants, it is wise to be sure that those types are represented in the sample. If a program is offered in different communities, the credibility of an evaluation is enhanced when evaluators take special care to be sure that participants from a variety of program locations are included. For example, an evaluation of a reading program in a large urban school system could be based on schools whose students represent different ethnic and economic backgrounds rather than just taking a random sample of students from all schools.

To minimize expenses and to complete an evaluation on time, it is often necessary to sample participants rather than observe an entire population. The strategy of selecting a representative sample and devoting resources to obtaining data from as many from that sample as possible produces a more representative sample than a halfhearted attempt to include the whole population would yield. For example, a 75% response from a random sample of 100 would provide more valid information than a 15% response from a sample of 1,000 even though the second sample would be twice as big as the first. Observers untrained in social science research methods frequently emphasize a sample's size rather than its representativeness; this misperception may require the evaluator to teach a little sampling theory to stakeholders.

No matter how sampling is to be done, evaluators know that some people will not cooperate and some cannot. Thus, it is advisable to seek a larger sample than is essential. Identifying the characteristics of nonparticipants and those who drop out at various points in time is important. Both the attrition rate and the reasons for attrition have a bearing on the conclusions the data can support.

Some evaluation models would not employ a representative sample, but only include those selected in accordance with the focus of the model. For example, in a formative evaluation, using a special sample of successful participants might permit one to develop insights into what pattern of program participation is needed to achieve success. It is critical to keep the purpose of the evaluation in mind when planning such details. Using such a selective sample is probably unsuitable for a summative or a quality assurance evaluation. Even when it makes sense to have a focused sample, unintended consequences need to be considered. Examining only those smoking cessation clients who attended all 10 of the workshops offered may fit one purpose of observing those who received a full dose of the intervention. But it also likely includes the most motivated clients, and their experiences are unlikely to be similar to all clients.

Another crucial step in selecting a methodology is choosing a research design. The choice depends on such constraints as the purpose of the evaluation, the preferences of the stakeholders, the deadline when the project must be completed, and the funds available. Some evaluations simply require a series of interviews conducted by a knowledgeable evaluator, others require a creative examination of existing records, and still others require a design similar to those used in basic research involving random assignment to program and control groups. Evaluators should remain flexible, selecting the methodology that matches the needs of stakeholders rather than routinely using a favorite method. The evaluation project proceeds more smoothly when program managers, staff, and representatives of those providing financial support have participated in the selection of the methodology and understand the reasons for the approach ultimately chosen.

Using multiple criteria from different sources yields the best information for an evaluation. The program participants themselves are one of the most important and accessible data sources. They can describe their reactions to the program and their satisfaction with it, and they can describe their current behavior. People close to the participants, such as family members, may sometimes provide information, especially when the participants are very young or incapacitated. Program personnel and records are often essential. For some programs, community-level variables are the most relevant. Last, evaluators must not overlook their own observations; this point is expanded in Chapter 8 on qualitative methods.

The day-to-day mechanics of data collection usually involve an on-site coordinator who keeps track of program clients, personnel, and other individuals supplying data for the evaluation. Surveys must be administered at appropriate times, addresses need to be updated, and relationships must be maintained with program staff. Because data collection is critical, yet sometimes tedious, a responsible person should handle the task.

In quantitative evaluations, appropriate statistical analyses need to be completed. Using simple statistical procedures whenever possible is desirable because the findings will be presented to stakeholders who may not have a great deal of statistical expertise. Stakeholders ideally should be able to understand the interpretation of

the results; they should not merely be impressed by statistical sophistication. One approach is to use statistical analyses that are as powerful and complex as necessary, but to illustrate the findings using percentages or—even better—graphs. Some ways in which analyses of evaluations differ from those that are used in basic research are mentioned at several points in later chapters.

During the planning phase, careful thought must be given to how an evaluation is to be reported to the stakeholders. Often evaluators have erred in focusing on a single report. Instead, a plan for reporting throughout the evaluation process serves the interests of the stakeholders and provides evaluators with constructive feedback as the project progresses. Chapter 13 contains an extended description of the variety of reporting avenues evaluators use to maintain support for the evaluation and to encourage the use of findings.

Present a Written Proposal

After reviewing the literature and thinking through the various methodological considerations outlined above, the evaluators are ready to prepare a written proposal. The acceptance of a proposal indicates that evaluators, program personnel, and perhaps the funding agency have agreed on the nature and goals of the program, the type of evaluation wanted, the criteria and measures to be used, and the readiness of the program for evaluation. It is psychologically and practically important for the program personnel to understand the evaluation process, to feel comfortable with it, and even to look forward to using the information that will be obtained.

Formal contracts are required for external evaluators. Such contracts will specify the tasks that the evaluator is to complete and any tasks for which the program staff is responsible. Items that went into the estimate of the costs should be specified. Internal evaluators cannot always insist on formal written agreements; however, listing agreements in memos with copies to program directors and central administrators helps in ensuring that the various stakeholders hold similar expectations for the evaluation. Expectations for an evaluation can be renegotiated if necessary. Using written communications ought not to be interpreted as a lack of trust because honest misunderstandings occur when memories differ. Such differences have created problems after evaluations have been completed.

DYSFUNCTIONAL ATTITUDES MAKING AN EVALUATION CHALLENGING

Political and emotional factors can undermine an evaluation project. Effective evaluators seek to identify these factors, discuss conflicts, and reassure stakeholders that the program evaluation is planned to serve their needs. Some concerns simply represent misunderstandings of program evaluation. Other concerns may reflect real conflicts within the organization that the evaluation has brought to the surface.

Our Program Is Exemplary

Most people seem to think that they are better than average in lots of ways (Epley & Dunning, 2000) even though on the average that cannot be. In a similar way program planners are generally enthusiastic and confident about the effects of their work; indeed, many expect their new program to have dramatic, positive results. Mendez and Warner (2000) showed that a goal published by the U.S. Public Health Service concerning the percentage reduction in adult smoking rates by 2010 was "virtually unattainable" because it was too ambitious. Program planners and staff might feel betrayed when evaluators find programs to be less than perfect. Difficulties are especially likely when a new program is expected to improve on a reasonably good existing program.

It may be necessary to assist the stakeholders in estimating the amount of improvement that is reasonable to expect. For example, if elementary school children are reading at appropriate levels, a new reading program cannot raise reading levels very much. The new reading program might be considered successful if reading levels increase just a little or if students like reading better. Furthermore, a small improvement experienced by many people could be very valuable (Cook & Shadish, 1994; Rosenthal, 1990).

There are other reasons why a dramatic effect seldom occurs. Whenever a group of people is compared with a group of program participants, the comparison group will usually be receiving some kind of service. Program participants rarely can be compared with a group that receives no service at all. Similarly, when

evaluators compare two versions of the same service, both of which are considered good, large differences between groups are unlikely. For example, if psychosomatic patients treated in a special unit are compared with those being treated by their own physicians but without the benefit of the special program, both groups should improve because both are receiving treatment for their immediate problems.

An Evaluation Will Offend the Staff

Evaluators recognize that no one likes to be graded, especially if those grades become public knowledge. In recent years, however, all sorts of published accountability information has been gathered and summarized. Even patient ratings of physicians are available on the Internet. It is critical that evaluators recognize that making program staff defensive will not contribute to a useful evaluation. Furthermore, evaluators need to remind people that program evaluation is not personnel evaluation. The evaluators focus on the *program*, not individuals who work in the program's activities. Having made this assurance, it may be critical to note that the idea of accountability and transparency has swept government and nonprofit organizations in the last decade.

An Evaluation Will Inhibit Innovation

Staff members may worry that evaluation will interfere with innovation by inhibiting them from experimenting with new techniques. In both process and outcome evaluations, the staff may believe that after beginning an evaluation no variation in the program is permitted until data collection is complete. This is partially true because major structural changes in the program could alter the essential goals or nature of the program. In such a case the evaluation would need to be revised. However, this does not mean that clinicians or program personnel cannot be flexible in the day-to-day operation of the program within reasonable boundaries. Every program has some flexibility; evaluation will not limit that. However, it is wise not to attempt to conduct an outcome evaluation of a program that is just getting started; major changes can occur as staff members become clearer about their objectives.

Our Program Will Be Terminated

Although a negative evaluation seldom leads to a program's cancellation (Cook & Shadish, 1994), it is possible that an evaluation could result in the curtailment or elimination of a program when findings show that a given approach simply is not working as expected. However, before sponsors can eliminate a program designed to meet a specific problem, they are ordinarily under some pressure to decide what to put in its place. Therefore, an unfavorable evaluation will more likely result in the refinement of a program than its elimination.

Early in the planning phase, effective evaluators will try to have program personnel view them as partners who share the goal of providing quality education, service, or care. Evaluators can assist program personnel in meeting accountability standards that may be a condition for continued funding. One practice that allays some anxiety is to promise the program managers that they will see the final draft of the report and be asked for their suggestions and clarifications (see Chapter 13). The comments of program administrators are appended to many reports done by evaluation and audit offices of local and federal governments. Of course, evaluators cannot eliminate anxiety completely when a program does not have the full support of central administrators or when there are valid concerns about the quality of the program.

The Information Will Be Misused

Besides the fear that the program will be canceled, there might be some concern that information gained about the performance of the staff may be misused. Even competent clinicians, administrators, teachers, and other personnel are concerned about merit reviews, future promotions, and career advancement. It is absolutely necessary to keep formative program evaluations distinct from assessments conducted to reward the more effective staff members. This point has been emphasized by writers in education, medical care, and industry. However, staff members realize that they probably cannot control the information about the

program that will be available to administrators of the organization housing the program.

Evaluation Methods Are Less Sensitive Than Perceptions of Staff

Service personnel rightly feel that their observations are valuable sources of ideas both for improving the functioning of a program and for evaluating its effects. They may feel that the evaluators' questionnaires, complicated research designs, and statistical techniques are less sensitive than their own observations and evaluations. At times they are right.

Although the subjective evaluations of the staff can be biased (see Dawes, 1994), the ideas of program personnel are a very valuable source of evaluation data. Evaluations can be improved by both quantitative and qualitative data gathered from many sources. The staff's subjective observations will be very important when the data are being interpreted. The ideal strategy is to integrate the findings from both methodologies as described in Chapter 8.

Evaluators gain the confidence of managers and staff not only by recognizing this problem but also by articulating this awareness in such a manner that program staff members are reassured that the richness of human services has been appropriately respected (Taut & Alkin, 2003). Early in the planning phase, program personnel can be assured that the evaluation will not begin until they have had the opportunity to review the evaluation proposal carefully and are confident that their concerns have been addressed properly.

Evaluation Drains Program Resources

The concerns described so far are focused on various aspects of evaluation but not on the concept of evaluation itself. Some objections to evaluation strike at the very idea of program evaluation. The staff may charge that program evaluation drains money and energy away from direct service (Botcheva, White, & Huffman, 2003). As the statement stands, it is true. However, a more critical issue is whether evaluation can improve service in the long run. The alternative to spending money on evaluation is to risk spending money on services that are of unknown value. Today it would be hard to find a program funded either by government agencies or by private foundations that is not required to evaluate its services. Those who are not convinced by the accountability argument may be persuaded by a more pragmatic one: evaluations, if done well, may help in attracting more support and resources into the program. Furthermore, a favorable evaluation could lead to additional settings adopting an effective program.

Evaluations Have Little Impact

Some critics point out that evaluations frequently have had very little impact on programs. There is a good deal of validity to this objection; evaluators have often been frustrated by seeing their reports set aside. However, evaluators should remember that when evaluations are used, their results will be only one of many factors behind decisions. Because evaluators work in a complex organizational context, the results of their work must be timely and relevant to decision making. Well-designed and carefully executed studies are valuable only when they address the issues that are important to the organization. When evaluators show how evaluation is relevant to pending decisions, they raise the odds of the evaluation being used.

The Effect of These Attitudes

Stakeholders who fear evaluations of their work might attempt to determine the methods to be used, urging evaluators to use marginally valid methods that are likely to produce a desired result. Ideally, discussions among stakeholders and evaluators will reveal that better procedures are available and should be used. Open discussion ought to reduce or resolve unreasonable fears; however, evaluators ought not to ignore the threatening nature of evaluation. At times, an organizationally powerful stakeholder may seek to control an internal evaluation. One way for an internal evaluator to resist that influence is to seek outside support in the form of a review panel.

CASE STUDY 2

Change 4 Good

For over 10 years, Regina Langhout, PhD, a professor at the University of California at Santa Cruz, along with a number of her graduate students (including Jesica Fernández) and undergraduate students, have engaged in a Participatory Action Research (PAR) project called "Change 4 Good." One aspect of the participatory nature of the project is that the goals and therefore the details were established by the various collaborators, the university researchers, and the fourth- and fifth-grade students in a nearby elementary school during the early stages—see especially the Meeting Needs section. In this case, "Students set a goal to increase the school's connection to themselves and their families" (Langhout & Fernandex, 2015, p. 198), with one major component being the creation of a mural that would present an image of the school that was better connected to students' stories.

This case particularly illustrates two relevant points. First, many aspects of evaluations are affected by the evaluation model or models chosen—in this case, the Empowerment and Capacity Building models. One clear example of the Empowerment model is that the project includes evaluations by the elementary students as well as by themselves. A number of other examples appear in the following sections. Second, along those lines, many good evaluations and their reports show important differences with other evaluation reports. For example, this evaluation included much more extensive comments about stakeholders, specifically using that term, than many other evaluations. Readers will likely notice other points as well.

Meeting Needs

A creative method, photovoice, was used to identify needs; this involved students taking pictures in response to a prompt about their hopes and then building on those pictures. The nature of the need identified fits the project and evaluation approaches—the students focused on a system problem, that "the school was not connected to them and their families, and therefore, they did not have a sense of belonging" (Langhout & Fernandez, 2015, p. 198), rather than on an individual problem such as some students not being engaged enough in the school.

Implementation

The article documents many points in detail about the actual implementation. These elements illustrate aspects of a formative evaluation—insights are used to make adjustments to the program to address problems or weaknesses. For example, the project involved creation of a mural at the school, but when students evaluated the project, they "determined that their peers did not feel very connected to the first mural" (Langhout & Fernandez, 2015, p. 203), so they decided to work on a second mural, making sure that students felt more connected to it.

Stakeholders

Consistent with the principles of empowerment evaluation, the report not only described the many relevant stakeholders, it also detailed multiple efforts to engage various groups of stakeholders in both the project and its evaluation.

Side Effects

In some ways, the broad nature of PAR and an Empowerment/Capacity Building evaluation means that most results fit more into the realm of desired outcomes than side effects. Still, a positive side effect was that involvement in the project provided the setting for children to demonstrate capabilities and performance that was less obvious than established academic assessment. A negative side effect was that some of the children found the difficulties of creating systemic change very discouraging.

Outcomes

Major points such as the creation of two murals, a book about one of the murals, and a documentary about the other are clearly documented outcomes. Fitting the Capacity Building purpose, they noted that the students adjusted their approach to collecting questionnaires and then were successful in reaching their target number, which illustrated the outcome of having the participants gain evaluation skills.

Nuances

Looking more at understanding various aspects of the events than simply identifying some mechanisms responsible for specific outcomes, the students looked at different responses to the mural by younger compared with older students, and students as compared with teachers. For example, they noted that the students liked elements of the mural that illustrated struggles more than teachers liked them, drawing the conclusion that students felt the images of struggles were particularly relevant to their own experiences.

Sources: Kohfeldt, D., Bowen, A. R., & Langhout, R. D., *Revista Puertorriqueña de Psicología*, 27, 2016; R. Langhout (personal communication, September 21, 2017); Langhout, R. D., & Fernandez, J. S., In D. M. Fetterman, S. J. Kaftarian, & A. Wandersman [Eds.], *Empowerment evaluation: Knowledge and tools for self-assessment, evaluation capacity building, and accountability*, Thousand Oaks, Sage, 2015.

EVALUATOR PROFILE 2

Regina Langhout, PhD—Scholar-Activist

Regina Langhout, PhD, is a community psychologist and social justice activist, or scholar-activist. A scholar-activist often works and advocates with those who are marginalized in various ways. As one example of her thoroughgoing and passionate commitment, she strives to work with all others in respectful, transparent, and engaging ways, from the elementary-school children in Change 4 Good to the school staff and administrators to her university students.

Given that a primary goal for her project and for the evaluation was Empowerment, it is not surprising that one focus was the children's growing ability to make sense of complex information and to act based on their understandings. An important part of that focus involved helping students learn not only how to run analyses such as t-tests with their data but also how to understand how the tests worked and what the results meant, along with taking responsibility to collect and manage the data. In keeping with the goals of Capacity Building, these elements meant that the students participated substantially in the evaluation rather than simply providing the raw data for her to analyze and use. The meta-message that she trusted the students would not have been lost on them, even if they might not have used such terms. These kinds of priorities are evident throughout her work and writings.

But Regina has also been deliberate in working to become even more engaged with as many collaborators and stakeholders as possible, noting various challenges when balancing expectations and values of multiple stakeholders. Specifically, she notes that one approach to such authentic work with school staff can be described as being a "Critical Friend"—a way to describe the challenging balance of being supportive but also being as honest as possible about weaknesses as well as strengths. And her lab meetings with her graduate students include time to acknowledge and address students' and her own life contexts from physical health to other stressors and celebrations as well as the focused business of research projects.

Although she does not identify primarily as a program evaluator, evaluation is clearly one critically important tool and approach that serves her commitment to scientific rigor while working for justice. And those long established in the field as well as those just entering it will find inspiration in her example of paying attention to her heart and to the passions and the values that sustain us (see Langhout, 2015).

Sources: Langhout, R. D., *American Journal of Community Psychology*, 55, 2015; R. Langhout (personal communication, September 21, 2017).

SUMMARY AND PREVIEW

Careful planning of an evaluation project serves to get an evaluation off to a good start. Note that the steps are suggested to help evaluators respond to the needs of the people most affected by the evaluation. New evaluators may find the basic checklist in Figure 2.3 to be a helpful guide for their initial projects. Responsive evaluators who cover the areas comprehensively have fewer problems with the fears outlined in the second part of this chapter than do evaluators who seem to ignore stakeholder needs.

The next chapter focuses on a central concern—specifying the implementation and outcome criteria of wisely planned and successful programs. A thoughtless selection of criteria will negate efforts to conduct a useful evaluation.

	Steps
_____	Learn the main points about the program—what it does and who it serves.
_____	Learn who are the relevant stakeholders and meet with appropriate representatives.
_____	Determine the parameters of the evaluation—who wants it, why, when, and with what focus?
_____	Review relevant literature regarding the program, the population served, and other relevant factors.
_____	Choose one or more appropriate models of evaluation.
_____	Plan the basic method of data collection, analysis, and process of determining conclusion.
_____	Develop a written proposal of the above elements.

Figure 2.3 Basic checklist of program evaluation steps.

STUDY QUESTIONS

1 Apply the evaluation planning steps to the hypothetical sexual assault prevention program mentioned at the beginning of Chapter 1. Make a list of what needs to be done if you were to contemplate evaluating such a program.

2 What models of evaluation would you think are most appropriate for the program to prevent sexual assault on campus?

3 One can think of a college course as a program. (In fact, a 3-hour class requires more time and energy than many social service programs.) The elements of the program consist of (a) reading assignments, (b) class discussions, (c) lectures, and (d) projects such as term papers. The outcomes that university and future employers value include (a) the mastery of content; (b) the development of information management skills such as writing, analysis of arguments, and synthesis of information; and (c) the development of meta-skills such as organization, self-discipline, and maintaining good interpersonal relationships. What are some needed intermediate steps if the elements of the program are to lead to desired outcome skills?

4 Imagine that you are part of an evaluation team in the institutional research office of a large university. The chairperson of the department of psychology asks you to evaluate a graduate program called "Community Psychology." List the stakeholders. What are the first questions to ask the chairperson? What models of evaluation seem most appropriate to such an evaluation?

5 What models of evaluation are implied if an administrator asks an evaluation office to conduct an evaluation of a program specifying that the program manager and staff are not to be aware that the evaluation is being done?

ADDITIONAL RESOURCE

Card, J. J., Greeno, C., & Peterson, J. L. (1992). Planning an evaluation and estimating its cost. *Evaluation & the Health Professions*, 15, 75–89.

When asked to estimate how much to charge for conducting a program evaluation, most new evaluators have little to go on. When pressed, most people will greatly underestimate how much time is

necessary to complete even a very modest project. Although this is an older resource, these authors suggest how long it is likely to take to complete various steps in the process. Evaluators making a cost estimate need to remember that they ought to be paid for the initial discussion and the preparation of the proposal itself. There are many sites on the Internet referring to costs (just enter "program evaluation cost" into a search engine); however, it was hard to find one that provided estimates of the time needed as this source provides.

3 Developing and Using a Theory of the Program

What is the best measure of a good intercollegiate athletic program? Some would say the proportion of games won; others would say the number of tickets sold; others, the devotion of the alumni to the team; others, the success of the student athletes after graduation; and still others, the favorable media attention attracted to the university through its teams. Some college presidents and faculty have argued that the criteria of a good sports program have shifted too much toward winning and away from the academic and life successes of the athletes after graduation. In fact, at some well-known schools, only small percentages of basketball and football players graduate. For the few who become professional athletes, failing to complete college may be a small matter, but for the others, not graduating limits their future vocational opportunities. Some people argue that this is irrelevant; basketball and football proceeds earn enough to support the entire athletic program, and, since no one compels the students to participate, the academic administration should not disrupt a good thing.

Differences about the desired criteria also lead to varied personal choices. If we test-drive a car, we use degree of comfort, ease of operation, clear vision of the road, perhaps what our friends think is cool, and other criteria, even though we have never written those standards down. When people disagree over the desirability of different automobiles, it is often because they are using different criteria to make their choices. For some people, an attractive exterior design is more important than a reputation for reliability; for parents with several children, a larger rear seat would be more important than it is for single people holding their first jobs.

Whether we think about them or not, we do have criteria that permit us to make distinctions and to form preferences as we make our choices of foods, friends, churches, automobiles, and politicians. When conducting an evaluation of a program or a product, we need to develop or select criteria and standards in a fashion that is far more explicit than we need for daily life, and those criteria are best when they are related to a theory appropriate to the program. Without the development of clear, appropriate criteria and specific standards, we might never be able to agree on the value of a counseling program for students who fear math, food stamps for poor families, or training for people who have lost jobs. As mentioned in Chapter 1, the traditional model of evaluation had been a subjective assessment by the people offering or funding services. Such informal assessments are no longer acceptable, especially when programs are paid for with public funds. This chapter describes the need to develop a program theory and then how to use that theory or a related model to choose the criteria that are to be measured and that will indicate successful implementation and desirable outcomes.

DEVELOPING A PROGRAM THEORY

Disagreements over criteria as illustrated above often may simply reveal disagreements over the purpose of a consumer choice or a program. Whenever people develop service or intervention programs, assumptions about the causes of the problems and the best ways to change problem behavior are made. Unfortunately, these assumptions are often implicit (Cook, 2000; DeFriese, 1990; Lipsey, 1993; Posavac, 1995; Posavac, Sinacore, Brotherton, Helford, & Turpin, 1985). When there is no explicit statement of the theory behind the development of an intervention or when there is no conceptual framework linking the interventions to the projected outcomes, it is hard to agree on the criteria indicating program effectiveness, conduct an effective evaluation, or improve the intervention (Pion, Cordray, & Anderson, 1993).

Why a Program Theory Is Helpful

Initially, many evaluations were carried out without an explicit statement of the theory behind the program

design. Nearly 30 years ago, DeFriese (1990) lamented that a sizable portion of the descriptions of proposed health-related treatment programs did not include credible descriptions of how the interventions are supposed to work. The value of articulating the logic of programs offered in community mental health agencies has been demonstrated (Yampolskaya, Nesman, Hernandez, & Koch, 2004). Increasingly, theoretical foundations are expected in evaluations (Rog, 2015) as it has become clear that program design is less effective without theory and evaluations are less informative when carried out in a conceptual vacuum, although there is clearly room for further improvement in actual practice (Munter, Cobb, & Shekell, 2016).

Specifying the theory behind the program provides assistance for planners, staff members, people responsible for obtaining funding, as well as for evaluators. In a comment on social welfare programs Etzioni noted "we know that it is extraordinarily difficult to change habits, personality traits, culture, and social institutions" (1994, p. 14). Yet governments and agencies frequently propose policies that are based on the assumption that interventions can motivate people to make major changes in their lifestyles in short periods of time. We act as though threatening shoplifters with jail will keep them from taking clothes from stores; telling patients with diabetes to lose weight will lead them to do so; showing teenagers how to recycle soda cans will stop them from discarding cans on the beach. Information is important; without knowledge of adaptive behavior people cannot change. Yet, there is ample evidence that such information alone is very weak compared to the major influences in life, such as peer pressure, family practices, media models, and simple convenience. Across many kinds of activities, it is clear that people differ in how ready they are to acknowledge problems and to take some action to change (Prochaska & Norcross, 2001). Sometimes evaluators when working with a program planning committee can lead the planners to clarify their assumptions and implicit theories and perhaps refocus their plans. Perhaps the plans should focus on more realistic objectives rather than spreading program funds too thin by attempting too much.

A second value of thinking theoretically is that we can identify the intermediate results of a program rather than just the final outcome. We are also reminded that different people might respond to different forms of a program. Teenagers might respond if a music or sports figure endorsed a program; on the other hand, business managers might want to see a financial approach to a social concern. Before participating, most people need to learn how to visualize themselves carrying out the recommended behaviors. A neighbor was overheard arguing with her husband about a newly announced community recycling program. Citizens were to place cans, bottles, and newspapers into orange-colored containers on Friday mornings. She was not enthused and asked her husband, "Are you going to sort through the garbage every night?" For decades she had put apple peels, the morning newspaper, coffee grounds, and empty soda cans into the kitchen garbage basket. Although keeping recyclable items separate from other garbage may have seemed to be a small matter to planners, the idea of separating items before putting them into the "garbage" was initially novel for some people. Publicity about a recycling program should include illustrations of how to participate in convenient ways. Describing the processes for people to follow in making changes in their lives can lead to more effective programs.

A third benefit of having a program theory is that theory helps us to know where to look in conducting an evaluation. Several decades ago, Lipsey and Pollard (1989) remarked that adopting "the basic two-step" in program planning and evaluation would improve common practice. Planners can specify (a) what should happen right after participation in a program as a step toward change and (b) a more ultimate outcome behavior that reflects a goal of the program. Evaluators would then develop methods to measure both. Learning about success or failure while the program is being provided is more informative than learning years later that the final outcome was or was not achieved. Particular attention would be paid to why the program did not lead to the intermediate result; perhaps additional resources were needed or the original theory was not valid. If the intermediate outcome was achieved but the final outcome was not, one possibility is that strong nonprogram influences were limiting the achievement of the desired ultimate outcomes, while another possibility is that the theory is wrong. Collecting information that could indicate which possibility is correct would follow from the theory and would be extremely useful to the program.

How to Develop a Program Theory

It is easier to agree that a program theory should be developed than to develop one. There are a number of

approaches to developing a program theory and differences of opinion over the detail required in preparing a model of the program. Not only do some argue for more complicated and complete models (Renger & Hurley, 2006), Gates and Dyson (2017) noted that recent approaches to understanding causality mean that evaluators have a higher bar to clear. Still, even simpler models are better than none (W. K. Kellogg Foundation, 2004).

In any case, good evaluators talk first with the staff of the program. At times the staff will have a fairly clear idea of how the activities making up the program affect the participants and how the intermediate stages lead to the final desired outcomes. Often, however, staff members are surprisingly vague. The rationale for some interventions may turn out to be as simple as "we have always done it this way," and there may be no clear expectation of what initial steps lead to the final goals. In such cases, evaluators then turn to the research literature on similar programs. The theoretical underpinnings of similar programs that are producing good outcomes can be examined for relevant insights.

The research literature contains two kinds of material that might prove helpful. First, evaluations of similar programs might describe program theories. For example, peer leadership might lead to favorable outcomes more consistently in antismoking campaigns in junior high schools than do teacher-led classes (see Evans & Raines, 1990). Since junior high children are especially sensitive to the attitudes of peers, program planners might try to use peers in any program that attempts to influence adolescents (Turner, 1999). In evaluations of energy conservation programs, evaluators have noted that electricity use is not easy for people to monitor; few people ever look at their electric meters and, when they do, it is hard to relate the spinning disk to energy use. On the basis of these observations, programs have been developed to provide new ways to give people feedback about their use of power in terms that are understandable (Seligman & Finegan, 1990).

Examining basic research is a second way to develop program theory. Social support is often believed to be related to health and positive adjustment. Approaches to help people obtain social support are discussed in social and clinical psychology research studies (e.g., Glasgow et al., 1997). Some of those ideas might be appropriated to enrich the theory of some social service programs. Unfortunately, writers of basic research often provide only brief descriptions of the independent variable, that is, the treatment, thus making it difficult to apply their work to other settings ("The trouble with dependent variables," 1990).

Throughout this process evaluators must remember that they cannot just pick a theoretical model for the staff. Instead, evaluators can learn what model or models are implicit in the program and work with staff to examine the relevant elements. This is an iterative process and may take some time. It is extremely important that evaluators learn what is intended to happen to the program completers, how they are to change initially, what program experiences are to lead to those changes, and what activities the program must contain. Without specifying those steps, evaluators cannot choose what to observe and no one will know what the information gathered showed about the program.

A particularly good, classic example of the development of program theory was provided by Cook and Devine (1982), who described the expected processes whereby a psychoeducational intervention would help postsurgery patients recover more quickly and experience fewer side effects. Figure 3.1 has been adapted from Cook and Devine to illustrate some of the processes that connect the nurses' teaching to the desired outcome, having patients able to leave the hospital sooner without ill effects. Note that there are numerous criteria that may be observed by someone evaluating the program; it is wise to include many intermediate variables, both those believed to be relevant and some likely alternatives, as well as the final outcome criteria of program success.

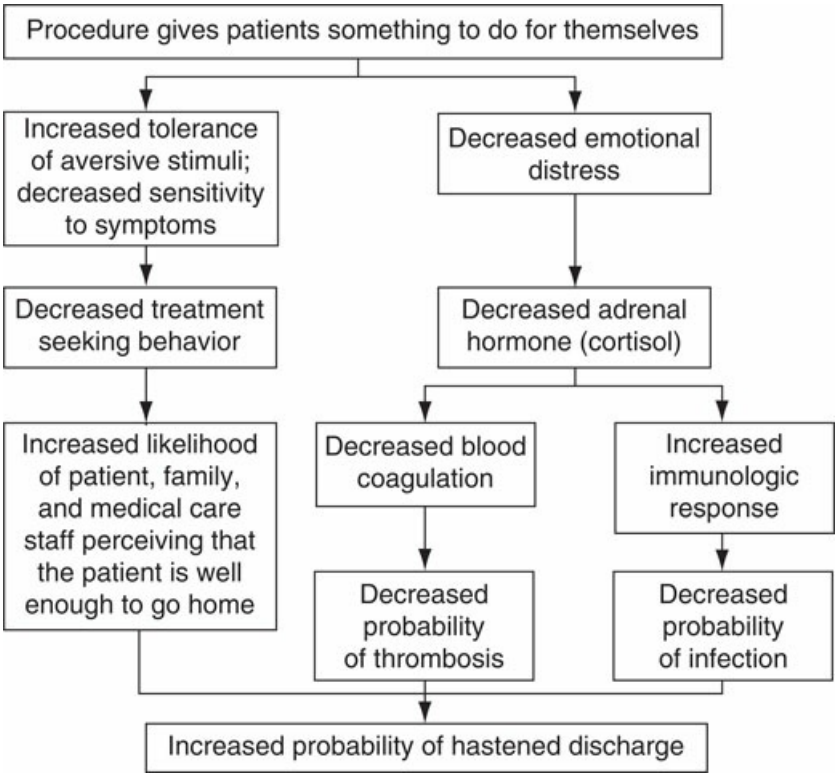


Figure 3.1 Illustration of an impact model showing the intermediate steps that are expected to be observed between the intervention and the desired outcomes. (Illustration prepared on the basis of the findings presented by Cook, T. D., & Devine, E. C. Trying to discover explanatory processes through meta-analysis. Paper presented at the National Meeting of the American Educational Research Association. March, New York, 1982.)

Implausible Program Theories

In the process of selecting the criteria of program success sometimes it becomes clear that program theories are implausible. This is not to suggest that a great proportion of program theories are implausible, but some are, and wise evaluators don't dismiss the idea that the conception of the program that they have been asked to evaluate might be based on implausible assumptions. Owners of swamp land were selling alligator hides to shoe manufacturers at a rate that led naturalists to worry about the disappearance of alligators in Florida. To protect alligators, the Florida state legislators passed a law forbidding the sale of alligator hides. This seemed like a straightforward way to permit alligators to thrive in Florida. Figure 3.2 is a diagram of the theoretical model underlying this law.

The outcome of the policy could have been evaluated by estimating the number of hides sold in spite of the ban. A comparison to the sales in years prior to the ban might well have shown that the new law successfully reduced the number of hides available. However, the real objective was not to have fewer hides sold; the real objective was to have alligators thriving in Florida. The legislators held an unstated assumption: it was assumed that landowners who were unable to sell alligator hides would maintain their land as alligator habitat. If that happened, the goal of the legislators would have been achieved; however, what happened was that landowners sought alternative uses of their land. By draining the land they were able to develop new farmland. When they did, the amount of habitat for alligators was reduced thereby negating the desired outcome of the legislation (Sieber, 1981). Detecting implausible program assumptions is more likely when theoretical models are developed and presented in a concise manner.

Policy	Immediate Objective	[Unstated Assumption]	Stated Desired Outcome
		[Landowners will maintain land as alligator habitat.]	Alligators will continue to thrive.
Forbid the sale of alligator hides	Reduce the number of alligators killed for their hides	**[Plausible Reaction to the Ban]** [Landowners will seek an alternative use of the land, such as draining it to develop farms.]	**Plausible Outcom** Alligators will have less habitat and will not thrive.

Figure 3.2 Sometimes programs are planned without a carefully thought-through impact model; when that happens there may be unstated assumptions in the model. At times unstated assumptions are implausible. When that happens the program can have an effect that was not imagined when the program was designed and implemented. (Model prepared from information in Sieber, S. D. *Fatal remedies: The ironies of social intervention*. Plenum, New York, 1981.)

EVALUATION QUESTIONS FLOW FROM PROGRAM THEORY

Since criteria and standards are chosen for specific programs, it is impossible to list the criteria that evaluators might use in conducting specific program evaluations. However, a number of general evaluation questions are central to program evaluation; most evaluations would include some of these questions. The value of dealing with these questions and the costs of poor choices of criteria are illustrated in the following sections.

Does the Program or Plan Match the Values of the Stakeholders?

Educational and social service programs are designed to fulfill purposes that are laced with values. For example, what should be included in sex education classes for junior high students? Should the government provide funding for abortions? Should welfare recipients be required to perform public service work? These questions cannot be answered by listing facts.

Evaluators provide a valuable service when they help stakeholders clarify their values and assumptions. The changes to the U.S. federal income tax code at the end of 2017 provided a clear opportunity to see the range of values and assumptions held by different people. Although most people would prefer they themselves pay less in taxes rather than more, most also believe that people with greater incomes should pay a somewhat greater percentage in taxes because they tend to use more resources than average. But what percentage should people at what income levels pay? The maximum rate was lowered in an attempt to prompt greater investment in businesses leading to more jobs. The effects of the changes in tax rates are hard to pinpoint, but the effort would be worthwhile to help in deciding whether the rate change was wise. The data will not affect fundamental value disagreements. However, identifying the value disagreements permits gathering data that speak to the central issues rather than to those merely peripheral to the arguments.

Turning to a different example, public housing programs are designed to make adequate housing available to poor people at affordable costs. Most people would agree that this goal matches the values of society. However, the way public housing policies had been implemented in large cities led to the clustering of poor people, often in high-rise buildings that neither foster a sense of community nor provide safe settings for residents (see "High-rise brought low at last," 1998). An evaluation of public housing ideally would include information on the conflict between the values on which public housing policies were based and the results that have occurred. While a program evaluation cannot tell anyone what values to hold, the demonstration of the discrepancy between what occurred and what was hoped for could be an important motivator for improvement.

Does the Program or Plan Match the Needs of the People to Be Served?

The unmet needs of some segment of the population usually form the basis for the development of a program. The most useful program evaluations compare the unmet needs of people with the services available through the program. Unemployed people need skills that lead to employment. An excellent training program for a skill that is not in demand would not fulfill a need. Programs cannot be evaluated in isolation from the community characteristics and the people being served.

In order to verify that a plan for a new program will meet the needs of the people to be served, evaluators

conduct a needs assessment (Witkin & Altschuld, 1995). An introduction to the methodology of the assessment of need is provided in Chapter 6. But some examination of need for the program is essential in most if not all program evaluations. Since program resources are always limited, changes in the relevant needs for various programs over time may suggest changes in the resources devoted to those programs.

Does the Program as Implemented Fulfill the Plans?

Although the issue of monitoring implementation is the focus of Chapter 7, it should be noted here that there are at least two aspects of implementation. One is the match between the planned program and what is actually being done, and Chapter 7 provides important details about evaluating that match. The second aspect is more closely related to program theory, however. It is one thing to claim that a program is based on a given theory, but demonstrating that the actual practices follow the theory is a different matter. This section addresses various aspects of determining whether the program matches the plan.

Failure to Implement the Theory of the Program

Evaluators first examine a program plan to learn if the expected mechanism assumed by the theory is valid, because integrating theory and practice is much more complex than just applying one to the other (Leviton, 2014; Schwandt, 2014). Observers provide numerous examples of situations in which the theory failed. Antabuse is supposed to make alcoholics ill if they consume alcohol. One method is to implant Antabuse surgically. In an early evaluation of the procedure, blood tests of only 8 of 31 alcoholics who were to be getting Antabuse by surgical implant showed the presence of a therapeutic level of medication (Malcolm, Madden, & Williams, 1974). A low-income population proved too hard to reach so a publicly funded family planning center served nearby college students (Rossi, 1978). Shopping mall cholesterol tests have been found to be exceedingly unreliable and, thus, of little use to people wanting to know whether they needed medical attention (*Cholesterol screening*, 1990). A quantitative measure of fidelity to program plans can be developed (Mowbray, Bybee, Holter, & Lewandowski, 2006).

Programs That Cannot Be Implemented as Planned

A program cannot be offered as planned when the population to whom the program was directed rejects the service. The community mental health system was proposed to serve chronically ill psychiatric patients in community settings rather than keeping them confined to large state mental hospitals. It was believed that antipsychotic medication would permit such patients to live independently or with family members and that community mental health centers would provide supportive care and medications. Unfortunately, many discharged patients rejected the treatment. Since people could not be compelled to go to the centers and since funding was not as generous as expected, the program could not be implemented as envisioned. Instead, many mentally ill people ended up as homeless in the streets and under the bridges of urban areas.

Sometimes the theory underlying a program is not appropriate to conditions under which it is to be implemented. A curriculum to teach mathematics in elementary schools (dubbed the "New Math") was well-planned, mathematically sophisticated, and evaluated favorably during development. However, the teachers who used it during development were better trained than the teachers who used it after widespread implementation. Accustomed to drilling children in multiplication tables, elementary teachers were expected to teach set theory. In addition, the new terminology made it impossible for most parents to offer their children any help. After a few frustrating years the new curriculum was discarded (Mashaal, 2006). An objectives-based evaluation that focused only on the degree to which the children learned the concepts covered could not detect a major reason why the New Math curriculum was ineffective: many elementary school teachers dislike and fear mathematics.

Resistance to the Program Can Make Thorough Implementation Impossible

People seeking to change the economic systems in Eastern Europe and the former USSR met with great

resistance from people who were used to the older, centrally planned systems. Furthermore, many people preferred guaranteed employment and artificially maintained low food prices over the promises of better economic conditions in a more free economy that were uncertain. An example closer to home occurs when college deans ask faculty members to spend more time doing one-on-one academic counseling with undergraduates and to report on the counseling done. Faculty members who feel pressure to make progress with their research or who may be more interested in graduate education have found ways to feign compliance with a dean's request while not changing actual behavior, perhaps by describing brief hallway contacts as academic counseling.

Do the Outcomes Achieved Measure Up to the Goals?

Evaluations of operating programs usually examine at least some results of the programs. Developers of innovative curricula would include measures of student achievement, and planners of an advertising campaign often depend on sales information. Even programmers of religious events deal with outcomes when they report the number of people who say that important behavioral changes took place as a result of participation.

Seeking Information about the Levels of Outcome Expected

Some outcomes can be easily compared with what would be expected without the program. Education and health measures are among the most well developed for setting specific goals. Norms for academic skills are available for standardized achievement tests, and normal ranges are well known for many laboratory tests used by physicians. The outcome of on-the-job training programs could be specified in terms of the trainees' skill in carrying out the tasks needed by the employer. However, even tests with norms that have been validly developed can still be misused. Critics of the use of standardized achievement tests to evaluate school districts (Cannell, 1987; Shepard, 1990) suggest that when schools are repeatedly evaluated using standard tests, teachers come to tailor their teaching to the test and may even teach specific aspects of the test to their students. Linn (2000) showed that when a new test is used in a school district, scores initially fall, but then rise each following year as teachers learn about the test.

When programs are in areas for which extensive norms do not exist, evaluators are often tempted to permit program staff members to specify the objectives of the program. Although evaluators need the contributions of staff stakeholders in all phases of evaluations, experienced evaluators frequently discover that staff members are *too optimistic* about the level of success that will be observed. When one can find completed evaluations of similar programs, the outcomes achieved in those programs can provide information on what might be achieved by the program to be evaluated. The more closely an evaluated program matches the program to be evaluated, the more relevant the previous findings will be. Such information provides some rationale for the levels of outcome specified in the objectives.

Deciding on the Meaning of Outcome Achieved

Deciding whether the outcome observed is good or only marginal depends on an estimate of what should have been achieved and what specific outcomes mean. For example, after a year of reading instruction, almost all third graders can read better than when they started school in September; however, school boards and parents will not be satisfied with just any level of improvement. The improvement should be commensurate with nine months of instruction in third grade. Thus, when evaluators and stakeholders specify goals and objectives, attention must be paid to how much improvement they expect.

When evaluators first begin to work with stakeholders to develop statements of outcome objectives, sometimes they borrow the style of writing hypotheses that they learned in statistics classes. In statistical analyses the alternative hypothesis is often presented as the mean of the program group exceeding the mean of the control group, H_a: $M_p > M_c$. In most social science research, however, the amount of difference between the groups is seldom specified. In contrast, merely showing that changes occurred as in social science research is not sufficient in evaluation. It is more useful to list the actual level of achievement desired for students, patients, and trainees. If a product is to be evaluated, the performance needed is specified. If a medical information system is developed for a large medical practice, just saying that the new system would be *faster*

than the old system is not very helpful. Instead, a minimum performance level should be specified. If the new system is not at least that fast, then the system is not working as well as planned. Evaluators trained in social science research methods often have difficulty committing themselves to a specific goal because basic research in the social sciences is not carried out that way. There are some alternative approaches to statistics for evaluators (see E. J. Posavac, 1998, and Chapter 13).

Evaluating without a Theoretical Model

The ease with which staff, government auditors, and the general public fall into the trap of skipping the step of model development suggest that the issue needs to be reemphasized. The Auditor General of Illinois criticized *Parents Too Soon*, a state-funded program whose objectives included reducing the rate of teenage pregnancy (Karwath, 1990). In defense, the program's administrator described the difficulty of evaluating a prevention program, and mentioned the reduced number of live births to Illinois teenagers between 1983 and 1988. (By using these figures, note that the administrator had adopted a nontheory-based evaluation.) Since the number of Illinois teenagers dropped between those years and since the numbers of abortions were not known, the decreased birth rate was not seen as support for the program. Furthermore, there are so many influences on the rate of teenage pregnancy that skipping the development of a program model placed insurmountable limitations on the interpretation of any findings. If the program had specified how the program was to work, implementation goals and intermediate goals, and the final, bottom-line outcome goal, the administrator may well have been in a better position to respond to the criticism.

Although the idea of integrating theory and practice is a compelling and valuable goal, the human tendency to use established patterns called heuristics to process information and reach conclusions means that we tend to do things in the same way we always have. Recognizing these habits and working to increase the flexibility of our thinking is thus an important strategy to improve matters (Patton, 2014). Along those lines, empowerment evaluation approaches and the associated method of Getting to Outcomes provides a structured way for organizations to build more systematic and rigorous reflection on their ongoing program and results (Wandersman, Alia, Cook, Hsu, & Ramaswamy, 2016). Reflecting systematically on one's program and the theoretical reasons for its elements is one way to address some of these problems (Leviton, 2014).

CAUTIONS IN CHOOSING EVALUATION CRITERIA

Once a program theory has been developed, it is possible to begin to specify the criteria of successful implementation and desired outcomes that will permit us to carry out useful program evaluations. Many evaluators have emphasized the importance of sound research design (as done in later chapters); however, just as a chain is only as strong as its weakest link, an unreflective selection of what to measure can lead to a failed evaluation just as surely as an inappropriate research design can. Ill-considered choices of standards could make it impossible to draw any conclusions from an evaluation even when the evaluation was well planned in other ways.

Criteria Must Reflect a Program's Purposes

Without very careful planning and a thorough understanding of a program, it is quite easy to select criteria that do not reflect its purposes. Many observers came to believe that the first evaluation of Head Start (Cicarelli, Cooper, & Granger, 1969), a popular, federally funded preschool program for children of poor families, was mortally wounded by the decision to use measures of the improvement of intellectual skills as the primary criteria of success. Instead, Lazar (1981) argued that the most important aspect of Head Start was the involvement of parents in the education of their children. It was hoped that the involvement of low-income parents, many of whom typically participate only minimally in the education of their children, would continue as their children grew. Thus, Head Start was hoped to lead to increases in the likelihood that low-income parents would encourage the development of their children, help with doing homework, and make sure their children were in school. If more parents adopted such practices, such activities could have long-term positive effects on the children, effects far more important than the specific knowledge that the children gained in Head Start classes (Leik & Chalkley, 1990).

What should the criteria for success be for a recycling program? Since it is necessary for residents to separate their recyclable materials from garbage, it might seem reasonable to measure the amount of recyclable material collected as an outcome criterion. But there have been reports of materials being collected, but then simply dumped in the regular garbage landfill. The real criterion of success is that the materials be reused in some way. A recycling program is actually quite complex. For such an effort to achieve desirable outcomes, the cooperation of residents is needed, along with an efficient collection system that meshes with sorting facilities that in turn distribute materials to appropriate manufacturers who find customers for their products. Criteria of success may look valid yet not measure success in achieving the ultimate objectives of the program.

The time when an evaluation criterion is measured could be chosen in a way that fails to reflect the program's purposes. A program could have positive immediate or short-term effects, but a marginal long-term one. A physical fitness program can be quite effective in improving muscle tone without having an effect on program participants' physical condition or health 10 years later. On the other hand, some college alumni remark how helpful a faculty member was for their development even though the faculty member was perceived as critical and overly difficult when they were in school. Here, the long-term effect seems more favorable than the short-term one. Evaluators seek to make observations at times which correctly reflect the objectives of the program and the needs of the participants.

Criteria Must Be under the Staff's Control

Evaluators can expect considerable resistance to an evaluation when program staff members feel that their program will be judged on criteria that they cannot affect (Torres, 1994). For example, employees in manufacturing firms may feel that their effectiveness is limited by the quality of the equipment they must use. Such limits need to be considered in evaluating whether they are doing as good a job as possible. Some college teachers object to the use of tests of student skills to evaluate teaching effectiveness because they fear that student achievement levels reflect student effort at least as much as faculty skills and effort. In addition, students enter college with different levels of developed skills; thus, the more well-prepared students graduate with better skills. Furthermore, if students must hold jobs while attending classes, their achievements will on the average be lower than that of others. Those whose success depends on the performance of others—as is the case for teachers, counselors, coaches, or managers—face a more uncertain situation than do people who are more directly in control of the results of their efforts. The challenge to evaluators is to find ways to identify measures and criteria that reflect the efforts and skills of the staff and are agreed to be central to the program's purpose.

Stakeholders Must Participate in Selecting Evaluation Criteria

The criteria for a specific evaluation are selected in close consultation between the evaluator and the stakeholders involved with the project (Renger & Hurley, 2006). If the stakeholders do not accept the criteria selected, even when they are appropriate, the evaluation cannot have the impact that it would otherwise have. This does not mean that evaluators simply use whatever criteria and standards stakeholders want. Stakeholders are seldom trained in evaluation methodology. When there is disagreement about the appropriate criteria and standards, evaluators spend many hours in dialogue with the stakeholders to assure that all parties agree on the standards discussed (McGarrell & Sabath, 1994).

The selection of criteria and standards also varies depending on what aspect of the program development process is to be studied. Programs go through a number of phases: proposal preparation, planning, initial implementation, actual operation, expansion, contraction, and accreditation and reaccreditation. Criteria and standards to evaluate a plan for a program will differ markedly from those used in a reaccreditation.

In addition, the criteria differ according to the type of program being evaluated. Programs in education, health, criminal justice, marketing, and training differ in their emphases, the relative power of different stakeholder groups, the traditional types of information found useful, and the financial support available for evaluation studies. These differences play a part in the development of standards of program success and are taken into consideration in designing program evaluations.

IDENTIFYING GOALS AND OBJECTIVES

In order to know how well we have achieved our goals, we need to know what we wanted to achieve. Some people set out on an automobile trip without a destination in mind. If driving is pleasurable and if simply a change of scenery is desired, having no destination is fine. But most people want to travel to some specific location. If stakeholders are unable to decide what a training or service program is supposed to achieve, there is little reason to begin the program or to evaluate it. A number of issues related to goals need to be addressed. Are there clear ideas about objectives? Are there disagreements about objectives? If so, are these disagreements mutually incompatible or could they complement each other?

How Much Agreement on Goals Is Needed?

Sometimes people are very unclear about what might be accomplished through a program (Cook, 2000; Mathison, 1994). Community groups may sense that there is a problem and wish to provide some service, but simply saying that we want people to be educated, healthy, and happy is not enough to develop a program to assist them. In other words, stating goals only in the most abstract terms may attract supporters, but it will provide little assistance in designing a program or in evaluating a program once it is under way. Abstract goals may be appropriate in the U.S. Constitution or in the mission statements of organizations, but will not do once people try to produce an operational program. There really is no reason to begin to plan if no one can describe goals any more specifically than saying that "we want to empower people," "we want to provide excellent medical care," or "we want excellence in education."

When stakeholders can describe the specific goals they hold, even if the goals are dissimilar, progress has been made and negotiations may begin. Progress will be stalled if the goals conflict with each other to the extent that achieving one goal makes it impossible to achieve the other; a family cannot vacation on Baja beaches and in the Canadian Rockies at the same time. Similarly, cancer care cannot make patients as comfortable as possible while using aggressive treatment with experimental medications. Planners can work with stakeholder groups to find mutually agreeable alternatives or to define situations in which one would use one set of goals rather than another. There might be ways to define a patient's condition that would determine when aggressive (but uncomfortable) treatment would be used versus situations when the goal would be simply to maintain comfort and to minimize pain while suspending intensive medical treatment.

Sometimes people have goals that are different but not incompatible. No college faculty is unified about what students are expected to achieve. Some professors expect detailed knowledge, others conceptual clarity. Some are concerned about the development of employable graduates; others are more concerned that students develop a sound philosophy of life. In most colleges, these different views exist side by side while students are exposed to a variety of professors and courses. In such a setting, mutual respect for differing points of view permits the organization to function.

Different Types of Goals

Once a program has been planned and started, there are different types of goals whose achievement can be evaluated. It is important to verify that the program has gotten underway as planned and to learn whether short-term and long-term outcome goals have been achieved.

Implementation Goals

All programs involve some level of activity before it is possible to achieve any outcomes. For example, equipment must be purchased and staff members must be hired and trained before any service is given. If these activities do not occur, there is no point in seeking to learn if the desired results have been achieved. Figure 3.3 includes a number of goals that the evaluator would need to examine in an evaluation of the effectiveness of a new Internet-based mathematics teaching system for middle school children. Implementation goals refer to goals that focus on the availability of the hardware and the quality of teacher training.

Although it may seem obvious to verify that the program exists before seeking to evaluate its effectiveness,

some stakeholders and evaluators have ignored this step. They seem to believe that if a program developer said that something would be done, then it is safe to assume that it has occurred as planned. One university vice president for research challenged the idea that the evaluation of implementation goals must be part of a program evaluation by asking, "Can't we assume that these are honorable people? Can't we assume that if they say they will do something, they will do it?" In fact the documentation of complete implementation is seldom a question verifying personal integrity. Instead, it is critical to learn if unforeseen problems made it impossible to carry out plans.

Examples of Different Types of Objectives for Internet-Based Math Instruction

Implementation Objectives

Teachers are to be trained to use the system and the curriculum chosen and must demonstrate proficiency according to the developer's standards by the beginning of the school year.

Schools are to have facilities to permit 15% of the children in a grade level to use the Internet system simultaneously by the beginning of the school year.

Intermediate Outcome Objectives

Over 70% of the children are to know how to use the system by November 1.

Children are expected to prefer this method of math instruction relative to the standard method used before.

Children are to complete at least 95% of the curriculum during the school year.

Learning Objective

The median achievement level in mathematics for children using the system is to exceed the median achievement of children using the standard instruction method in the school in the previous year by at least one tenth of a grade level.

Figure 3.3 The objectives specified for a program should include more than just the ultimate goal of the program. Implementation, intermediate, and achievement goals or objectives of a hypothetical program using an Internet-based instructional system for mathematics education for middle school children are illustrated.

A graduate student team sought to evaluate the effectiveness of a program in which faculty members volunteered to invite freshman, nonresident students into their homes for dinner in an effort to encourage the students to identify more closely with the college and, thus, be less likely to transfer to a different college at some time. Each faculty volunteer was given the names and addresses of 10–12 new students. The evaluators learned that 40% of the volunteer faculty members did not invite any students to their homes. Of the students invited, only 60% showed up. This meant that, at best, only 36% of this phase of the program had been implemented. The faculty members' intentions were not questioned. Scheduling problems may have come up; for some faculty members the task may have been more difficult than they had imagined; others may have misunderstood what they agreed to do; and perhaps some volunteered before checking with their spouses. Without including an implementation phase in this evaluation, the evaluators would have been unaware that the program was only one-third as strong as planned.

In Figure 3.3 access to the system and the training of teachers are the focus of the intermediate goals. If the achievement of students is to improve, students must be able to use the system and teachers must understand it.

Intermediate Goals

Intermediate goals refer to things that are expected to occur as a result of the proper implementation of the program, but that do not constitute the final goals intended for the program. For example, all students should

have textbooks (an implementation goal) and they should read the assignments on time (an intermediate goal), but the actual criteria of success go beyond reading the assignments. Intermediate objectives in Figure 3.3 deal with student acceptance and use of the novel method of learning.

Outcome Goals

Even if the Internet system in Figure 3.3 is implemented and used, it is still necessary to learn whether math achievement improves. We cannot be satisfied with simply verifying that the teaching system is in place: we must also examine its outcome goals. The most ideal finding would be that the program was implemented as planned, that the intermediate goals were achieved, and that the outcomes were favorable. Even then we would still not have conclusive proof that the program *caused* the achievement of the valued outcomes. However, finding that good results followed faithful implementation is compatible with believing the program was responsible.

There are other possible patterns of findings. Suppose that the intermediate goals were met (the hardware was in place, the teachers knew how to use it, and the students spent time with the computers) but student achievement did not improve. That would support those who felt that the more interactive method for learning math did not serve these middle school children well. It should be noted that for this example it was assumed that the evidence was good that the hypothetical Internet-based learning system was effective. Evaluators must never forget that although some groups may assume or claim that a program has good outcomes, it might not actually be effective (Weiss, Murphy-Graham, Petrosini, & Gandhi, 2008).

Evaluation Questions That Apply to All Programs

Critics argue that those who use objectives-based program evaluations limit the focus of evaluations to the explicitly stated goals of the program being evaluated, and devote little attention toward other factors (Scriven, 1991; Stufflebeam, 2001). If other aspects of a program were not examined, then objectives-based evaluation would have major limitations. One way to avoid this problem is to recognize that there are many goals that apply to all programs even though such goals are not listed during program planning. Program planners seldom need to state that people are to be treated with respect, that school children are not to be treated in ways that foster dependence, or that people are not to be discriminated against on the basis of race or sex. Since criteria such as these do not appear in lists of program goals, some evaluators have not treated these issues in program evaluations.

All programs also have the seldom-stated goal of no negative side effects or at least sufficiently minor ones. Evaluators should strive to become sufficiently familiar with the program and participants so that they can recognize any serious negative side effects. The negative side effects of Prozac, the antidepressant medication, became apparent (Glenmullen, 2000) in spite of enthusiastic advocacy of its widespread use (Kramer, 1993). At times evaluators have avoided personal exposure to the program in a misguided quest for objectivity, depending on surveys and records to provide evaluation data. Evaluators who do not become extremely familiar with the program, its staff, and its customers, clients, students, or patients run the risk of overlooking side effects.

Is the Program Accepted?

Good ideas are seldom adopted immediately. Mosteller (1981) described the astonishing lag in applying the sixteenth-century finding that citrus fruits (vitamin C, of course) would protect seamen from scurvy. On some long voyages, over 60% of the crew members died. One might have expected officials to immediately implement a policy that would have reduced this staggering loss of life, but 198 years elapsed between the first of several published descriptions of how to protect crews from scurvy and the adoption of the knowledge by the British Navy in 1795.

People for whom a program is designed may also reject it. There is a long line of research that clearly shows that physically ill patients often reject or ignore the treatment prescribed for their conditions (Posavac et al., 1985). Attempts to understand noncompliance with medical treatment suggest that better communication between physicians and patients promotes compliance. Some patients misunderstand the

recommended treatment. It is especially hard for some patients with chronic illnesses such as diabetes to maintain the motivation to adopt permanent lifestyle changes. Involving supportive family members is often an effective approach.

Many program managers request participants to complete surveys at the end of a service to measure the participants' satisfaction with the program. We believe that such surveys have a role to play in program evaluation, but it is important to recognize that such surveys cannot provide a sufficient means to assess the degree to which the program was effective. Some people are very satisfied with worthless treatments because they feel respected and valued by the people offering the service. However, a program that is disliked and rejected may not have a chance to be effective because participants who dislike a program or its staff are unlikely to participate fully. Marketing research firms focus on learning the preferences of potential customers and then matching products to those preferences. Although human service fields cannot base service design only on what students, clients, and patients think they need, program staff members who are aware of client preferences have a better chance of designing and offering a service that will be accepted and utilized.

Are the Resources Devoted to the Program Being Expended Appropriately?

There are many ways in which program costs can be used in program evaluations. Elementary procedures that relate cost to program outcomes are introduced in Chapter 11. The various evaluation questions that can be treated are reviewed here.

Using Program Costs in the Planning Phase

When a government program is developed to provide a service to all who qualify, it is difficult to project its ultimate cost because it is difficult to estimate how many people will apply. However, when a specific program such as a new curriculum or a counseling service for a set number of participants is being planned, the costs can be fairly well estimated. Sometimes it may be possible to project the value of the results. For example, job training should result in better jobs at salaries that can be estimated. Then, the taxes those people will pay and the amount of welfare assistance they will not use can be estimated. One could ask whether that result justifies the use of the amount of government funds needed to provide the program. If such a program succeeded in providing job skills and leading people to be self-sufficient, there may be additional benefits to which dollar values cannot be assigned.

A second issue in the planning phase concerns whether alternative uses of the funds would provide a more desired result. Since resources are limited, it is important that critical stakeholders agree that the services to be supported are needed more than other services that could be supported instead. Such decisions are based partially, not totally, on the basis of costs and the value of benefits.

Is Offering the Program Fair to All Stakeholders?

The use of ethnic origin to target educational aid is an attempt to give minorities a boost to overcome past discrimination. Many people endorse this policy (Dawes, 1994); however, some writers argue that such programs primarily assist children of middle-class minority parents who do not need the assistance (McWhorter, 2000; Steele, 1990). Evaluations of some programs aimed at increasing the skill levels of minorities have found that the programs actually increased the difference in skills between the children from low-income families and those from the middle class (Cook, et al., 1975), possibly because middle-class parents made special efforts to obtain those services for their children. If so, is it fair to continue to use public funds to support the program? A similar point can be made regarding "corporate welfare."

A second illustration concerns support for public education. Because American public education has been funded largely through local property taxes, wealthy communities and communities with large numbers of businesses and industries can spend far more per pupil than can poor communities without industrial firms. Some critics then argue that it would be more fair if school funds were dispersed on the state level so that all school districts can offer school programs of a similar quality. The evaluator cannot answer these questions of fairness using research methods. However, the use of standards of social justice in evaluation has been

encouraged by some writers (see Sirotnik, 1990) who argue that a completely objective evaluation methodology is neither possible nor desirable. It would be critical to be explicit about how such values are being used; evaluators with hidden agendas do not serve stakeholders well (Scriven, 1997a).

Is This the Way the Funds Are Supposed to Be Spent?

Another question about the use of program funding concerns whether the funds are being spent in a way that is compatible with the intention of the funding stakeholder. This question is related to the traditional accountability issue, which is designed to reduce the possibility of the misappropriation of funds. In the past, Congress and state legislators were often more interested in the question of whether funds had been spent as intended rather than the more difficult-to-answer questions of whether the funds had been spent wisely and whether the activities supported had met the needs that prompted the interest of the legislators.

Do the Outcomes Justify the Resources Spent?

Once the results of a program are known, it is possible to examine the program costs to determine whether the outcomes were worth what it cost to attain them. If it is possible to place a dollar value on the outcome, then we can ask if the return was worth more than the cost of the program. Business-based financial analysts do this all the time. If the expected return on a company's investment does not exceed the investment, then there is probably something better to do with the investment. If there really is nothing better to do with the money, then the business should be sold and the money distributed to the stockholders. When reasonable estimates of value are possible, this cost/benefit analysis is highly instructive. Although such analyses are becoming more common, some suggest that many outcomes require a more complex consideration of value that cannot be given a price tag (King, 2017).

Appropriate time frames also need to be considered. Sometimes U.S. businesses are criticized for demanding a return on investments too soon; it may be that sizable returns require a long-range perspective. In human service and educational fields it is quite hard to place a dollar value on the results of programs. In this case, evaluators seek to compare programs that are designed to affect similar behaviors, such as developing employment or reading skills, and to ask which program is the most efficient. Unless there are other constraints, the most efficient would be the best program to offer.

Has the Evaluation Plan Allowed for the Development of Criteria That Are Sensitive to Undesirable Side Effects?

If it were possible to predict specific side effects, program plans would be changed to reduce the chance of these negative outcomes. Since evaluators and managers expect some unanticipated outcomes to occur, observation procedures are planned to permit side effects to become apparent. This means that evaluators will allot time for visiting program sites and talking with representatives of all stakeholder groups. An organizational consultant remarked that although the management of a company hired him, he accepted a project only after it was agreed that the union would codirect the project. There is no point in carrying out an evaluation if the evaluator only hears one view of the strengths and weaknesses of a program or an organization.

An improvement-focused approach to evaluation includes attention to unplanned results, whether positive or, as is often the case, negative. Overlooking side effects is more likely when evaluators plan an evaluation on the basis of official program descriptions (rather than the program as implemented), obtain information from only one stakeholder (usually the manager), or carry out the evaluation at arm's length from the program. A professor of community nursing described how one evaluation team tried to evaluate a program designed to provide medical care to dispersed poor families in suburban areas. From the comfort and security of their offices they prepared a questionnaire based on the official objectives presented in the grant proposal and simply mailed the survey to a list of participants. These evaluators never visited any of the sign-up centers, never interviewed any participants, and never contacted any of the cooperating physicians. This mechanical application of objectives-based evaluation has contributed to negative reactions to evaluation in general and objectives-based evaluation in particular.

When evaluators believe that they have detected negative side effects, they will share their observations with stakeholders to obtain confirmation of their views or additional information to explain the observations. More careful observation procedures can then be developed to examine the unanticipated problem. Finding negative side effects is not done to discredit programs, but to assist in improving programs. Perhaps the program theory can be improved (Fitzpatrick, 2002). Sometimes it helps to have an outside observer call attention to overlooked problems, but calling attention to a problem is merely the first step; for effective evaluators, a harder and more important step involves helping to develop plans for improvements.

SOME PRACTICAL LIMITATIONS IN SELECTING EVALUATION CRITERIA

Textbook authors have the freedom to describe ideal practices without being limited by the practical realities of specific evaluations. Three important restrictions on the selection of criteria are the evaluation budget, time constraints, and the degree to which various criteria are accepted by the stakeholders.

Evaluation Budget

Evaluation is not free; evaluators and their staffs need to be paid, surveys need to be duplicated and mailed, costs for phones and computers add up quickly, and getting around to various program sites can be expensive. Evaluators estimate the amount of time they will need to do a credible job in light of the stakeholders' needs. It is usually necessary to negotiate the parameters of the evaluation since few stakeholders know how much time is involved in completing a valid evaluation. Since the funds for an evaluation will be fairly inflexible, the focus of the evaluation is often adjusted during these negotiations. It is better to carry out a modest evaluation whose findings can be trusted than to plan an ambitious project that can only be done poorly given the resources available.

Time Available for the Project

Since evaluation is an applied discipline, the use of findings is often tied into budget cycles, academic years, or the meeting schedules of governmental bodies. Some years ago the federal government commissioned a study of an innovative approach to welfare. By the time the demonstration project and evaluation were completed, the political climate had changed and Congress was no longer interested in the original idea or the evaluation. Often, projects are of a much shorter duration than expensive multisite demonstration projections costing tens of millions of dollars, but the principle is the same: There is a time when information can be used and a time after which it is no longer relevant and will not contribute to decisions (Sonnichsen, 2000). Before evaluators accept a project or begin to plan seriously, they ask about the project deadline. If the deadline and the stakeholders' evaluation needs are incompatible, the evaluator must bow out or the project must be renegotiated. There simply is no reason to begin a project that cannot be completed when it is needed.

Criteria That Are Credible to the Stakeholders

Evaluators seek to collect valid data in ways that permit valid interpretations; in addition, it is crucial that stakeholders accept those interpretations. While planning the evaluation, evaluators ascertain that the stakeholders have agreed that the jointly selected criteria fit the goals of the program and are appropriate for the participants in the program. We talk about stakeholders "buying into" the criteria: it is helpful to have agreements in writing. The stakeholders need not sign a statement of acceptance; however, the evaluator can keep careful notes during meetings as an evaluation is planned. After each meeting, copies of these notes can be distributed to critical stakeholders. At the beginning of each meeting, the agreements summarized in the notes can be reviewed to be sure that stakeholders still agree on the objectives of the program and the focus of the evaluation.

CASE STUDY 3

"CU Flourish": A Medical Resident Wellness Initiative Based on Positive Psychology "PERMA" Theory

This case illustrates several important points. First, it provides another good example of the specific elements of a real evaluation. In this instance, it has the added benefit of being an evaluation in process, so more information will follow on a yearly basis in the eResources. Second, it illustrates how internal evaluators can be particularly appropriate (see especially the Improvement Focus section). Third, although many of the other cases were, in fact, conducted by more than one evaluator, this case explicitly notes the team effort that was involved. As you will see, the Evaluator Profile that follows presents the four primary team members rather than the story of just one.

Meeting Needs

Upon completion of medical school, physicians embark upon an intensive period of specialty internship and residency training that ranges in duration from three to more than 10 years. Training program requirements are demanding and are characterized by especially long days and weeks that require intensive periods of demonstration of clinical proficiency, academic achievement, and professional and personal stamina. Medical residency programs can take an appreciative toll on physician well-being. Taken together these demands are often reflected in high rates of burnout, depression, and in recent years, increased incidence of suicide (Dyrbye, et al., 2014). Accordingly, medical residency programs are now required to include wellness programs and/or initiatives in accredited residency training program curricula. The University of Colorado (CU) School of Medicine Resilience Program responded to these needs with the development and pilot implementation of a Resident Resilience curriculum entitled "CU Flourish."

Implementation

The development and implementation of CU Flourish has been accomplished in three phases:

Phase 1: Identification of theoretical and empirical underpinnings of curriculum components. The CU Flourish team adopted a unifying theory for the program's goal of addressing the broad range of specific needs: Seligman's (2011) Positive Psychology "PERMA" Theory of Well-Being, which is comprised of five building blocks that foster human flourishing—Positive Emotion (P), Engagement (E), Relationships (R), Meaning (M), and Accomplishment (A) (Seligman, 2011). They then developed four main program topics: "Finding Your Values Compass," "Mindfulness for Healthcare Professionals," "True(er) Grit," and "Putting Joy and Gratitude to Work"; these collectively address the five building blocks of the PERMA theory.

Phase 2: Initial program component pilot among varied constituent groups across the CU School of Medicine. Each of the topics was presented to colleagues and supervisors in the school, and the team reviewed the materials and feedback extensively in preparation for a two-day retreat format with CU Anesthesiology Residents.

Phase 3: Pilot study and dissemination of evaluation findings. The retreat was conducted with an entering cohort of Anesthesiology Residents. Subsequently, a "Train the Trainer" session was held for other interested physicians, beginning the dissemination of a standardized program.

Stakeholders

The CU Flourish program team members are accountable to multiple and varied "horizontal" (colleagues in various departments/specialties) and "vertical" administrative stakeholders. The School of Medicine administrators also are interested in demonstrating to the medical residency accrediting body (ACGME) that the programs are actively addressing the wellness-related needs of residents of each specialty. The residents who benefitted from the program are, obviously, also important stakeholders.

Side Effects

The negative side effects most immediately felt by the team were time pressure, limited resources, and demand for program components taken out of context from the CU Flourish program. As is often the case, the promise of an effective program was met with considerable demand for the program and the "Train the Trainer" sessions quickly, straining available resources. In addition, other stakeholders wanted parts of the CU Flourish program without uniform agreement regarding maintaining program integrity, fidelity, or common modes of outcome assessment.

Positive side effects included increased program visibility and enhanced cohesiveness and productivity of the evaluation team.

Improvement Focus

It is true that internal evaluators can potentially risk overstating good results partly because they themselves are stakeholders. In this case, the multiple examples of seeking feedback from colleagues and supervisors as well as the residents who were served by the program, and the repeated revisions to the materials in response to the feedback, demonstrate a high level of commitment to improve the program. In fact, this is the other side of the internal evaluator "coin"—being stakeholders also means that they not only care about improving the program, they have "skin in the game," so their dedication to improvement can be particularly high. Further, this evaluation team demonstrated a meta-evaluative approach to improvement—not just examining the implementation and impact of the program, but considering both the results and the process of the evaluation itself, such as noting the problems of pressure on the evaluation team. Such reflective practices are particularly good examples for new evaluators to follow.

Outcomes

Outcome measures were administered via online surveys prior to and after the retreat along with more lengthy surveys conducted at the beginning and end of each academic year. Some key quantitative outcome indicators of the program are (1) positive-to-negative affect ratio (i.e., positivity or "flourishing" ratio), (2) mindfulness scores, (3) satisfaction in valued areas of living, and (4) burnout scores. At the time of this writing, only baseline (i.e., beginning of the academic year) data have been collected for each of three Anesthesiology Resident cohort groups. During the retreat, evaluation data were collected at the conclusion of each of the topic sections. Both objective ratings and subjective comments were collected. Notably, results suggested that the program and topics were well received and deemed to be useful,

and there were significant increases in pre- to post-retreat flourishing ratios. More extensive outcome survey data will be analyzed when both the beginning and end of year time point surveys have been completed for all resident cohort groups and posted in the eResources.

Nuances/Mechanisms

The nuances/mechanisms of the program lie in the underlying PERMA theory. The premise is that if the PERMA components are found in the topic sessions, these will function as the active mechanism to elicit change in our key outcome measures. For example, the "Mindfulness for Healthcare Professionals" topic session was consistently rated by attendees to have Positive emotion (P), Engagement (E), and Meaning (M) components. Attendees noted that although the "Finding Your Values Compass" session did not necessarily tap into Positive emotion, it did elicit Engagement (E), Relationships (R), Meaning (M), and a sense of Accomplishment (A).

Summary

As the chapter emphasized, programs that have no explicit or even implicit theory make it challenging for evaluators to determine some of the appropriate measures, especially potential mechanisms. This case presents a good example of the benefits that accompany clear theoretical foundations.

Sources: A. O. Beacham (personal communication, February 2, 2018); Dyrbye et al. (2014).

EVALUATOR PROFILES 3

Abbie O. Beacham, PhD; Alison Brainard, MD; Norah Janosy, MD; and Jennifer Reese, MD

The core CU Flourish team is comprised of three physicians and one clinical psychologist, who all contributed vital elements to the program. They asked to present profiles for each member because they have functioned in this project as one unit. These colleagues are each passionate about physician wellness and resilience, contributing countless hours in training, education, and research in this area. In addition, each has a unique history that has compelled her to pursue this work, and each one's story and perspective makes this work compelling for others. The following "thumbnail" profiles provide an introduction to them.

Abbie O'Ferrell Beacham, PhD, is an Associate Professor in the Departments of Psychiatry and Family Medicine and an Associate Director of the Resilience Program at the University of Colorado School of Medicine. She completed her doctoral training at the University of Louisville, Louisville, Kentucky, and her predoctoral internship in Clinical Health Psychology at the University of Florida Health Sciences Center, Gainesville, Florida. Her post-doctoral fellowship research and clinical work was in psycho-oncology and orofacial pain at the University of Kentucky Colleges of Medicine and Dentistry. She has extensive clinical experience working and training students in medical settings including primary care, psycho-oncology, and chronic pain management. Dr. Beacham's research and clinical work has integrated mindfulness and acceptance-based approaches to behavioral self-management in clinical medical and nonclinical populations and, most recently, resilience in healthcare professionals.

Norah Janosy, MD, is an Assistant Professor at the University of Colorado in the Department of Anesthesiology. She practices pediatric anesthesiology at the Children's Hospital of Colorado. She attended medical school at the University at Buffalo State University of New York School of Medicine & Biomedical Sciences, completed her residency at Oregon Health & Science University in Anesthesiology, and completed a fellowship at Oregon Health & Science University in Pediatric Anesthesiology. Dr. Janosy is the co-director of wellness for her department with her colleague, Dr. Alison Brainard. She is actively involved in the Resilience Council at the University of Colorado, as well as a founding committee member of the "Peer-to-Peer" Network. She has presented on physician wellness and resilience at numerous national conferences (Stanford Med-X, Colorado Review of Anesthesia for Surgicenters and Hospitals [CRASH], Society of Education in Anesthesia [SEA]). Physician wellness and resilience is a topic about which she is passionate, and she has devoted her academic career to this important topic.

Alison Brainard, MD, is an assistant professor at the University of Colorado School of Medicine Department of Anesthesiology. She attended medical school at Saint Louis University School of Medicine and completed her internship at the Thomas Jefferson University Program and her residency at the University of Pennsylvania Program where she was Chief Resident, Anesthesiology. Dr. Brainard is the co-director of wellness for her department with her colleague, Dr. Norah Janosy. She is actively involved in the University of Colorado Resilience Council, as well as a founding committee member of the "Peer-to-Peer" Network. She has presented on physician wellness and resilience at numerous national conferences (Stanford Med-X, Colorado Review of Anesthesia for Surgicenters and Hospitals [CRASH], Society of Education in Anesthesia [SEA]), and has been an invited guest speaker at Yale University. Physician wellness and resilience is a topic about which Dr. Brainard is passionate, and she has devoted her academic career to this important topic.

Jenny Reese, MD is an Associate Professor of Clinical Pediatrics, Section of Pediatric Hospital Medicine at the University of Colorado School of Medicine. She is the Interim Section Head and Medical Director of the Section of Pediatric Hospital Medicine. She also serves as the Inpatient Medical Director for Children's Hospital Colorado. Her clinical time is spent as a pediatric hospitalist, and her administrative duties include quality and process improvement and clinical leadership, as well as developing and promoting programs that support wellness and resilience for health care providers. In 2015, Dr. Reese formed the University of Colorado School of Medicine Resilience Program for Faculty, Residents and Fellows, and she serves as the director of this program. Her undergraduate and medical school were completed at University of Washington in Seattle, Washington and her pediatric internship and residency were completed at the University of Colorado School of Medicine and Children's Hospital Colorado.

SUMMARY AND PREVIEW

The importance of developing a theory of a program and selecting evaluation criteria wisely is hard to overemphasize. The criteria are the windows through which users of the evaluation see the program. If the windows distort the view, the program would be misrepresented, either favorably or unfavorably. Ideally, evaluators observe more than just the final outcome of programs that require the cooperation and effort of the participants, as do programs in educational, psychological, criminal justice, and other settings. Evaluators and their clients benefit by examining the processes whereby a service is thought to lead to the expected outcomes. Even a very simple program theory that specifies program activities, expected intermediate outcomes, and final outcomes greatly improves the evaluator's ability to understand the program.

The next chapter focuses on the specific steps that one takes in measuring the criteria selected. Although a textbook separates the selection of criteria from a discussion of measurement issues, evaluators consider whether criteria can be measured validly and reliably at the same time they are making choices of criteria to measure.

STUDY QUESTIONS

1 Consider a setting with which you are familiar: school, work, church, team, dormitory. Each has rules or procedures to meet certain objectives. Try to analyze the assumed impact model to learn how the procedures might lead to the objectives. For example, there are required courses or area requirements for most college curricula. How might these requirements lead to educated graduates? What intermediate outcomes need to occur? How much validity does your impact model seem to possess? Drawing an impact model often helps to detect implausible assumptions. Try to find a few implausible assumptions underlying a public policy.

2 This chapter suggested that different stakeholders have very different views of how intercollegiate athletes should be evaluated. Make a list of some public policies or programs that would be likely to be seen as having quite different purposes depending on the stakeholder group evaluating the policy. You might consider the views of different groups regarding traffic flow, casino gambling, or drinking laws.

3 Show how an evaluator might select criteria of program success that are easily measured, but that miss the central point of the program. If you have trouble starting, think about the criteria of success for a Little League baseball team or a volleyball coach in a middle school. Those should be easy; now work with the meaning of success for college teaching, counseling, law enforcement, or other important community activities.

4 A local health care center participated in a program to teach volunteer community residents how to use computers to find health information posted by the center on its Internet site. Some stakeholders postulated that this effort would result in fewer cases of low birth weight babies in the community. Is this a plausible result of such a program? What might be more plausible outcomes from such a program?

5 What are some values and some limitations to the contributions that potential participants can make to specifying the criteria of successful programs from which they might receive services?

ADDITIONAL RESOURCE

Mertens, D. M., & Wilson, A. T. (2012). *Program evaluation theory and practice*, New York, NY: Guilford.

This comprehensive work provides detailed information, examples, and guidelines for those learning the field as well as for seasoned evaluators. In particular, the second of four sections presents a framework of paradigms, theories, and branches to explain the range of models and approaches used by different evaluators as well as for different settings and needs.

4 Developing Measures of Implementation and Outcomes

After evaluators and stakeholders have agreed on criteria that would indicate successful implementation and outcome, evaluators face the task of developing methods to measure those criteria. Psychologists commonly gather data from individuals; sociologists gather data on a community or neighborhood level; and economists focus attention on still larger groups. These differences reflect the questions treated by these disciplines. Programs are designed to affect individuals (e.g., education or rehabilitation), communities (e.g., crime prevention, building improvement), or states or regions (e.g., clean air policies, economic stimulation). Depending on the program being evaluated, program evaluators gather data in ways similar to the methods used by psychologists, sociologists, business managers, or economists. Regardless of the focus of the program, each type of data has strengths and weaknesses; consequently, evaluators should use multiple sources and methods. One chapter cannot cover all approaches. Instead the most frequently used sources and methods as well as the characteristics of good measures of evaluation criteria are reviewed. Last, some measurement tools are presented to illustrate the principles.

SOURCES OF DATA FOR EVALUATION

A dilemma for evaluators is that the stakeholders who know the program best may be biased toward seeing program success, whereas those who have the least self-interest often know little about the program. This dilemma leads evaluators to seek data from a variety of sources using different types of measurement techniques.

Intended Beneficiaries of the Program

Educational, training, medical, and welfare programs that focus on individuals usually involve information from individual participants. In contrast, programs in economic development, preventive medicine, and criminal justice focus on community-level variables rather than on specific, identifiable individuals. With such programs, all members of a community are, in a sense, recipients of the service; consequently, data that reflect the state of the community will be most relevant.

Program Participants

Program participants can provide information that is not available from other sources. In some situations participants spend more time with program staff than managers who are responsible for assessing the competence of the staff. College students, for example, often know more about the quality of a professor's teaching than does a department head or a dean. Furthermore, only the program participants know what their reactions are to the program. For many programs, participants are the best source of information on their own condition. Interviews with deinstitutionalized chronically ill psychiatric patients showed that for many patients, discharge damaged their sense of well-being (Shadish, Orwin, Silber, & Bootzin, 1985). An advantage of participant-provided information is that it is often relatively inexpensive to gather and is at least as accurate as other assessment approaches for a variety of behavioral and emotional dimensions (Shrauger & Osberg, 1981). There are, however, surprising limitations to what people can report about themselves. For example, people have difficulty recalling what they ate the previous week and most people report eating

medium-sized servings even though the definitions of "medium" varied by 200% in different versions of the same survey ("On a Diet?" 1989). Schwarz and Oyserman (2001) have shown how small differences in the phrasing of questions can affect what people report about themselves or their surroundings.

It is valuable to recognize that most participants can provide good data on many objective aspects of a program but not on other aspects. General hospital patients usually know whether rooms are clean, whether nurses and resident physicians treat them politely, and how long they have to wait in radiology (see Brown & Adams, 1992). However, they cannot evaluate the choice of medication or the competence of their surgeons. Similarly, college students can report on whether a teacher returned tests promptly, held classes as scheduled, and lectured or led class discussions. There are few undergraduates who can evaluate the accuracy of information presented in lectures unless the professor is grossly incompetent.

Experienced evaluators do not forget that an evaluation is not a high priority for program participants because they have sought the services of the program in order to meet a need, not to provide data. Many people will share their views if they believe that their efforts can help improve services for others. However, few are so altruistic that they will struggle with poorly organized surveys. The participants who have the most favorable impressions of a program or facility are the most likely to cooperate with data collection. The 86 percent who responded to a lengthy survey on a chaplaincy internship were independently evaluated by their former supervisors as having performed better than the 14 percent who did not return the survey (Posavac, 1975). College alumni who returned a mailed survey had earned higher GPAs than those who did not (Hogan, 1985). The choice of measures can minimize some of these problems, although this limitation should also be noted in reports.

Participants will need to be assured that their answers to surveys or interviews will be treated confidentially. Many people do not understand the social scientist's disinterest in facts about individuals, nor do they understand the necessity of using group averages and proportions in evaluation reports. They are familiar with the case study approach often used in newspaper and television discussions of medical, correctional, counseling, and educational programs. At times, however, details about an individual can play a role in an evaluation report by giving vivid illustrations of very good or very bad program performance. Using case studies of teenagers, Love (1986) showed that troubled adolescents were unlikely to fit into the mental health and welfare systems of a large city. Although a case study cannot show the extent of a problem, it can show how specific program failures occur and can help policy makers visualize a need for program development or change. Success stories can help planners as well because the way participants succeeded may suggest how to improve the program for others (Brinkerhoff, 2003; Kibel, 1999). To honor standards of confidentiality, any information about individuals should only be used in ways that do not permit the individuals to be identified.

Artifacts

In some settings, evaluators can use things that program participants or community members produce. In schools, students write essays, term papers, theses, school newspapers, and yearbooks. These artifacts reflect the quality of the curriculum, the teaching staff, and, indirectly, the school administration. In vocational settings, it might be useful to track the quality of items made and the amount of material wasted. The success of morale-boosting efforts might be revealed by a greater use of coffee cups and caps marked with a company or university logo. Turning to unwanted artifacts, the need for a vandalism prevention program might be demonstrated by high rates of broken windows, unkempt yards, and doubly or triply bolted doors. Success of a community program might be indicated by a lower rate of such artifacts.

Community Indexes

Some programs are designed to improve community-level variables. Examples include a crime prevention program to increase citizen participation in an effort to reduce theft and a citizen-developed program to monitor housing code violations to increase or maintain the quality of housing in a neighborhood. Since such programs are delivered to a community, the outcomes must be measured differently from the ways outcomes intended to affect individuals are measured. Community records, such as high school graduation rates, crime reports, fire losses, or immunization rates, may be summarized.

A major difficulty with community-level indexes is that there are many steps between the program and the desired outcomes (Cook, Leviton, & Shadish, 1985). Many community events, government policies, and personal habits may have such strong effects on the behavior to be changed that even a well-conceived program may not create a detectable effect. Such influences are, of course, beyond the control of the program staff. An evaluator who measures community-level indexes for an outcome evaluation of, say, a media-based preventive health effort while ignoring the integrity of program implementation and intermediate outcomes may be left quite in the dark about the reasons for apparent success or failure.

Providers of Services

The staffs of programs have information that is crucial to program evaluations. Programs also maintain records that, when summarized, can provide important clues concerning a program's operation and the outcomes achieved.

Program Staff

Staff members have been trained and are expected to assess participants' needs and improvement. Also, staff members know how well a program is managed and run on a day-to-day basis. On the negative side, staff members expect to see improvement; after committing themselves to helping people improve their skills, health, or adjustment, it becomes easy to overlook program problems.

Beyond a concern over the expectations of staff, evaluators need to recognize that evaluating a program is in some ways an evaluation of the staff's performance. Few staff members unhesitatingly welcome an evaluation of the effectiveness of services they provide; in fact, most will have questions about how an evaluation is to be used. If they are worried, they may resist the evaluation. Some college professors try to bar students conducting course evaluations from entering classrooms, and some psychotherapists argue that checklists of symptoms and rating forms are not sensitive enough to detect subtle, yet crucial, changes in clients that only therapists can sense. Although it may be hard to counter all concerns, program managers can assure staff that program evaluations are designed to evaluate programs, not individuals, and evaluators can assure staff of confidentiality. Furthermore, evaluators can include staff members in a thorough discussion of the findings of the evaluation as part of the report preparation process; providing scheduled progress reports as well as a final report can assure staff members that they will receive feedback throughout the evaluation process (see Chapter 13).

Program Records

Evaluators often begin developing specific methods to assess the criteria by asking about a program's records (Hatry, 1994). Records should include client characteristics, types and extent of services provided, costs of services, and workloads of staff members. When the variables are objective, these records can be very reliable. When the variables are less objective, as is the case with diagnoses or comments on progress, evaluators must exercise caution in using the information.

There are several important advantages of using the program's archives: the measurement process cannot affect the program participant, there can be no participant loss due to refusal or inability to cooperate, and data collection is less costly than would be the case when gathering data in the field. However, the quality of records may be poor. For example, a jail-based drug rehabilitation treatment program had such poor records that attempts to match treatment information with other information about clients had to be abandoned (Lurigio & Swartz, 1994). Hospital discharge records to be used in an evaluation contained many coding errors (Green & Wintfeld, 1993). Furthermore, records can be deceptively altered in anticipation of an evaluation, as done by the Inspector General of the Pentagon (Margasak, 2001).

When working as an agent of the service provider, an evaluator has legal authority to use confidential records; however, the importance of maintaining confidentiality cannot be overstated. Program records often contain material which, if released, could harm a program participant or staff member. Even the accidental release of critical information could leave an evaluator open to legal challenge; if insufficient care is taken

when handling program or agency records during an evaluation, the evaluator would gain a very unprofessional image, would lose credibility, and could be denied further access to records.

Observers

Observers who are not part of a program often provide perspectives that neither participants nor staff can. Observers could be (a) experts who are familiar with a wide range of similar programs or the needs the program is designed to meet, (b) people especially trained to rate the variables that describe the process or outcomes of a program, (c) people with relationships with participants such as family members or work associates, or (d) the evaluators themselves.

Expert Observers

Very complicated programs with abstract goals are often evaluated by having a team of experts examine the program's structure and products (Averch, 1994), which matches the Expert Opinion model, described in the Chapter 2 eResources. Accreditations of hospitals, rehabilitation facilities, and colleges are handled by experts who examine much information and talk with staff and participants before rendering a judgment. A basic research program must be evaluated before it can be provided with financial support. It is hard to think of a way to evaluate a proposed research program that does not use expert opinion. Experts not involved in the program being evaluated may be the least biased sources of information. Expert opinions may be essential to clarify the meaning of quantitative information describing the program.

Trained Observers

Many programs can be evaluated using people who have been trained to rate the degree to which a program was implemented as planned and well-defined criteria of program success. A municipal program to repair and clean parks might be evaluated by training a team of employees to examine parks using checklists with objectively defined levels of maintenance (Greiner, 1994). An evaluation of treatment services for severely mentally disabled clients may utilize ratings of psychological function by trained observers using a social adjustment scale (Green & Jerrell, 1994).

Significant Others

If a program is designed to change behavior (as correction and counseling programs are) or to improve health or work skills, people who have contact with the participants outside of program settings often can provide important information on improvements or problems (e.g., Katz & Warren, 1998). With the exception of hospital patients or prisoners, staff members see participants for far fewer hours per week and in fewer settings compared to family members who see participants in natural situations. However, significant others also have biases; they want to see improvement in troubled family members. Evaluators seeking information from significant others would want to learn about specific changes in participants; general subjective opinions would be less useful.

When requesting information about participants from significant others, it is necessary to have the participants' permission beforehand. People differ in their willingness to have spouses or others provide personal information to an evaluation staff. Requests for cooperation can emphasize the potential to improve services for others. Care should be taken to avoid implying that good treatment for the participant is contingent on giving permission to contact a significant other or on providing information for an evaluation.

Evaluation Staff

In some situations information is gathered by people playing the role of clients (Turner & Zimmerman, 1994). The number of busy signals, time to answer telephone calls, and time on hold can indicate agency responsiveness to citizen needs (Jason & Liotta, 1982). The City of Chicago Consumer Services Department

placed calls for taxi service to learn whether companies were honoring a requirement to serve disabled riders; the department learned that 47% of such requests were ignored (Washburn, 2001). Some forms of qualitative evaluations are based on the thorough observations by an evaluator familiar with social science methods and intervention programs, but not with the specific program being evaluated. When a disinterested observer examines the program, traditional practices and assumptions made by the staff may be questioned (Guba & Lincoln, 1989). Fresh insights into side effects of the program may be detected. Both the goal-free and the naturalistic models of evaluations depend on information gathered by the evaluators themselves. Evaluators must be free to observe the program and must remain on guard for attempts by agency staff to guide evaluators only to what managers want the evaluators to see. Figure 4.1 summarizes the primary strengths and weaknesses of the various sources of information just introduced.

Sources of Information	Strengths	Weaknesses
Intended Beneficiaries of the Program		
Individual participants (surveys or interviews)	Can provide information on the operation of the program Possess unique information on reactions to the program Can demonstrate change	Are not experts on programs Could be biased to see improvement
Community residents (surveys or interviews)	Unique information on the visibility of programs planned to affect communities Not likely to be predisposed to see accomplishments	Are not experts on programs May be unaware of the program (which could be valuable information)
Providers of Services		
Program staff	Can provide information on the operation of the program from an insider's perspective	Likely to be favorably disposed to the program
Program records and documents	Unique information describing the program, staff, and participants No problem with refusal to respond	Sometimes have been changed to support the program May be incomplete
Observers		
Expert observers	Familiar with similar programs Not likely to be biased in favor of the program	Visits are brief Observers might be misled
Trained observers	Can focus on the issues of particular interest Not likely to be biased	Might be expensive to train and to maintain on the program sites
People important in the lives of the participants	Best source of information on the day-to-day behaviors of the participants	Could be biased to see improvement
Evaluation team members	Can provide targeted information Not likely to be biased	Likely to be an expensive form of data collection

Figure 4.1 Different sources of information used in program evaluations have different strengths and weaknesses. Using multiple sources is the best approach to obtain comprehensive information.

Which Sources Should Be Used?

The choice of data sources depends on the type of decision to be made based on the evaluation, the size of the program, the time available to conduct the evaluation, and the cost of obtaining data. If an agency needs information to satisfy a state auditor's office, it is likely that a summary of records, a sampling of participants'

attitudes, and a fiscal report will suffice. If a decision to expand (or eliminate) an expensive, controversial service is anticipated, a more thorough analysis of the needs and progress of the participants coupled with the judgments of outside experts may be essential. Such decisions must be made on the basis of an evaluation of considerable rigor and credibility.

An important kind of data source is the general area of social media—the online networks such as Facebook and Twitter that have seen explosive growth in recent years. Several aspects of social media mean that they are potentially transformative ways to gather information of many forms relevant to program evaluations. They not only can connect evaluators with substantially larger numbers of participants, but can target specific populations relevant to an evaluation, and costs are generally much lower than alternatives. Even compared to just a few years ago, their use is becoming much more common (for example, Goldman & Beaumont, 2017; Miguel et al., 2017; Price et al., 2017).

The advantages of greater numbers of people who fit specific populations that might have been difficult or even impossible to reach using more traditional methods is hard to overstate. As is well known, one important aspect of social media is the way they can facilitate connections among those with very specific commonalities even when there is great physical distance among members of these communities or when typical interactions would not help members identify one another. For example, most social media provide access to people throughout most of the world, and even groups that might avoid any public identification due to stigma, such as those with particular diseases, conditions, or other aspects of identity that experience prejudice and discrimination, can still connect with one another, especially with the various levels of anonymity that online connections afford. Trying to recruit people to respond to surveys or other methods of data collection from many of these groups previously involved enormous challenges, whereas the possibility of reaching substantial numbers of those with almost any particular focus is now greatly increased.

These advantages are not unmixed blessings, however. For one thing, although the potential to connect with specific samples seems great, it is not clear how representative these groups are of the desired populations nor how online behaviors such as following links match established sampling methods that form the basis for established forms of analysis (Gabielkov, Ramachandran, Chaintreau, & Legout, 2016). Another important area of uncertainty is how established systems of ethics such as basic points of informed consent translate to social media. On the one hand, it is fairly clear to almost everyone that interacting online is much more similar to actions in public than in private. To put it differently, people generally understand that they have no reasonable expectation of complete privacy when engaged online. But on the other hand, a number of traditional areas of concern regarding protections in research raise important questions, such as obtaining consent from parents or guardians for those under the age of consent or what responsibilities researchers have when learning about illicit activities (Henderson, Johnson, & Auld, 2013). Perhaps what is most clear at this time is that social media will continue to grow as an important avenue of data collection for program evaluation just as it is in other areas, but there are likely to be various new developments that are impossible to predict with certainty now.

Regardless of the reason for an evaluation, evaluators should strive to use multiple measures from more than one source (Shadish, 1993). In choosing multiple measures, it is especially important to select measures that are not likely to share the same biases. For example, a client's subjective assessment of the success of a program may be mirrored by the subjective feelings of the client's spouse. If subjective measures of a program's success are supplemented by objective measures, evaluators are less likely to obtain sets of data affected by the same biases. When a conclusion is supported from a variety of sources, it becomes more credible to stakeholders.

It is, however, possible that the different sources will not agree. Shipley (1976) conducted an evaluation of a program in which college students served as volunteer companions to discharged psychiatric patients. The volunteers, patients, and hospital staff all viewed the program in quite glowing terms. However, more objective measures of the patients' behavior (staff ratings and records of frequency and duration of rehospitalizations) revealed highly variable outcomes. Some patients apparently benefitted; others did not. The use of multiple measures led Shipley to a conclusion different from that drawn in evaluations of similar companion programs using only subjective ratings. In a similar vein, Sullivan and Snowden (1981) found that staff, clients, standard tests, and agency files did not agree about the nature of the clients' problems. Massey and Wu (1994) showed that patients, family members, and case managers in a mental health facility assessed

the patients' level of functioning differently. When evaluating publicly funded job-training programs, subjective satisfaction might not be highly correlated with objective outcomes, such as income level (Hougland, 1987). Evaluators must therefore be sensitive to the possibility that focusing on a single source of information can distort their conclusions.

Since evaluators frequently work with people trained in service delivery techniques, not in research methods, evaluators often are surprisingly free to choose the criteria of effectiveness and the analyses of those criteria. Some observers have gone so far as to say that evaluators can determine the findings of an evaluation before it is conducted through their choice of criteria and analytic methods (Berk, 1977; Berk & Rossi, 1976; Zigler & Trickett, 1978). Fairness to all stakeholders requires evaluators to examine their own attitudes as they design evaluations. The following section describes the most important issues to consider when deciding how to gather information for evaluations.

GATHERING INFORMATION WISELY

When making choices among possible data sources and approaches to measurement, evaluators strive to use techniques that will contribute to valid, sensitive, and credible evaluations. Following the practices discussed below will help achieve this goal.

Use Nonreactive Measures

Nonreactive measurement refers to procedures of gathering information that do not prompt the respondent to give particular answers. For example, whenever it seems obvious that certain answers are desired by interviewers, some respondents tend to give those answers. The previous section hinted at another problem of reactive measures: merely focusing attention on a variable can lead people to change their behavior. Even surveys can be reactive because they may prompt respondents to think about issues they may have never concerned themselves with before. Clearly when an evaluation leads to changes in the behavior of staff or clients, the evaluation does not accurately reflect the program in operation. Unless the evaluation depends solely on existing records, totally nonreactive measurement is an ideal seldom achieved, although care in designing evaluations can reduce reactivity (Webb, Campbell, Schwartz, Sechrest, & Grove, 1981).

Use Only Variables Relevant to the Evaluation

Because program evaluation is an applied discipline, the variables selected must be relevant to the specific informational needs of facility management, community representatives, those responsible for budgets, and other stakeholders. In planning an evaluation, evaluators seek to learn what decisions are pressing but not understood. Discussions with staff and managers often reveal the important variables. If no decisions can be affected by a variable, or if no program standards mention the variable, then it probably is not essential to the evaluation. Evaluators take care not to measure variables that some stakeholders might think would be "interesting" but that would not affect the operation of the program or agency; such variables raise the cost of an evaluation but do not contribute to its value. Sometimes evaluators expect that a variable would be seen as important once summarized for agency leaders; at those times experienced evaluators act in the role of educator. Furthermore, evaluators seek to be sure that they are sensitive to good and bad side effects that cannot be anticipated.

Use Reliable Measures

Reliability refers to the consistency of the information that a measure produces for the same person in the same situation (Murphy & Davidshofer, 2005). There are many ways to think about reliability. If different observers describing the same thing report similar levels of a variable, we say that the observation procedure is reliable. If different levels are reported, we say that the procedure is unreliable. Reliability is calculated using the correlation coefficient, *r*. If two observers agreed completely after making a number of observations, a correlation of the two sets of data would be 1.00, perfect reliability. To the extent that they do not agree with each other, the correlation would be smaller; 0.0 would indicate that knowing one observer's reports would tell

us nothing about the other observer's reports. Generally, reliability values of 0.7 or above indicate good enough consistency.

If an outcome measure is a 50-item test of knowledge gained from, say, a job training program, there is another way to estimate reliability. The scores that a group of trainees earned on half of the items (say, the odd-numbered items) can be correlated with the scores for the other half of the items; such a correlation is called a "split-half" reliability. Note that the focus of interest is still consistency just as in the first example concerning interobserver reliability. For split-half reliability the issue is interitem consistency, that is, the different items seem to be measuring the same thing. This form of reliability in general is called internal consistency and is typically reported as Cronbach's alpha, which is the average of all possible ways to divide items for split-half reliability. Statistical analysis programs make such calculations simple.

Reliability is higher when the measurement procedure is minimally affected by the passing moods of the observer or the individuals observed. Like validity, reliability is likely to be higher when observations are based on objective behaviors rather than on inferences. Reliability is also higher when the measurement instrument contains more items. You may well have a sense of the higher reliability produced by using many questions if you think of a classroom test. A single item forms a larger portion of a short test than it would of a long test. If you happened to forget the topic covered by a certain question, your score would be affected much more in the context of a short test than in the context of a long test. Because a single question can have a major effect, short tests are less reliable than long tests. The more reliable a measure is, the more likely it is to detect an effect of a program (Lipsey, 1990).

Evaluators trained in disciplines in which the individual assessment is important are familiar with the reliability of classroom, achievement, or job selection tests. It is crucial that evaluators recognize that using tests to evaluate individuals is markedly different from using tests to evaluate programs. Since evaluators focus attention on groups, not individuals, surveys and observational techniques that would not be reliable enough for individual assessment may be acceptable to estimate the mean score of a group. By using the standard deviation of a measure and its reliability, it is possible to calculate a range of values likely to include a person's true score. For example, assume a person scored 105 on an IQ test with a standard deviation of 15 and reliability of 0.95. The true scores of 68 percent of the people getting 105 will be between $105 \pm 15\sqrt{(1-0.95)}$, or 101.6 and 108.4. The value added and subtracted to the person's score is called the standard error of measurement and applies when measuring an individual's score (Murphy & Davidshofer, 2005). This range is called a confidence interval.

By contrast, when estimating a group's mean score, the standard error of the mean should be used to calculate confidence intervals instead of the standard error of measurement. Suppose a group of 100 students are selected at random from a school. Suppose further that they have an average of 105 on the same IQ test mentioned above. For a group, we work with the standard error of the mean, which, in this case, is $15/\sqrt{100}$, or 1.50. We can conclude that 68 percent of such groups have true mean scores between $105 \pm 1.50\sqrt{(1-0.95)}$, or 104.7 and 105.3. Compare this confidence interval with the one found for an individual in the paragraph above. We can be much more certain of the likely value of a group mean than the score of any one of the individuals in the group. Even a measure with a reliability of only 0.60 would yield an estimate that would be useful for program evaluation purposes because evaluators are concerned about groups, not individuals. Even short scales are often sufficiently reliable for evaluating the effects of college experiences when samples are large (Walvoord, 2004).

When observer ratings are used, evaluators also seek high interobserver reliability. An instrument that yields similar values when used by different people has more credibility than one that yields different values. Obviously, if the ratings are done in different settings—at home versus at school—a low interobserver reliability index may be observed due to real differences in the person's behavior between the settings themselves. Still, interobserver reliability can be increased through the development of a detailed instruction manual and careful training and periodic monitoring of observers.

Use Valid Measures

The instruments must validly measure the behaviors that the program is designed to change. These behaviors

may be the ultimate criterion behaviors (such as a healthy life, developed skills, employment) or more immediate behaviors believed to be useful in achieving the planned long-term outcomes (such as more consistent adherence to prescribed medical treatment, higher quality work, or improved leadership practices). In order to increase the credibility of the evaluation, evaluators make sure that staff members have approved the measures. But because program staff members are often inexperienced in social science methodology and cannot judge how feasible it is to measure a particular variable, evaluators do not depend on them to suggest all the measures and variables.

In general, the more a measurement tool focuses on objective behavior rather than on ill-defined or vague terms, the more likely it is to be valid. Objective behaviors (such as "late to work less often than once a week" or "speaks in groups") can be more validly measured than traits (such as "punctual" and "assertive"), regardless of the specific measurement approach chosen (Mager, 1997).

Even when the variables seem quite objective, information taken from official records may not be as valid as it first appears. Changing definitions and differing levels of care in preparing records can introduce changes in the meaning of variables that could influence an evaluation. At least twice Chicago's crime records appeared to reveal an increase in crime when better recordkeeping was begun by the police. After a well-respected person was installed as chief of police, thefts under $50 increased dramatically. Instead of ushering in a crime wave, the new chief simply insisted on better reporting procedures, and may have reclassified some crimes. Since homicides are usually recorded accurately and cannot be reclassified into a different category, homicide rates were used to determine that the crime rate did not increase when the new chief took over. In 1983 a similar increase was observed when recording standards were again strengthened ("Good News," 1983). After the renorming of the SAT, average SAT scores of college freshmen showed an increase between 1994 and 1995, a result solely due to a change in scaling by the Educational Testing Service ("Notebook," 1994). The same pattern can be seen in many other cases.

One other aspect of validity particularly relevant to evaluations is important to note. Although validity has often been considered on a general scale, the concept of accurately measuring the intended target can be applied in multiple ways. There has been growing recognition in recent years that just because a measure is valid for one population does not mean it must be valid for all populations. An obvious example is that a measure written in one language is not valid for use with people who read a different language. In fact, even translating the measure into another language is not necessarily sufficient because words in two languages that seem to refer to the same thing may mean different things to those who speak those languages.

Even within the same language, there are important cultural differences that must be considered when attempting to measure elements validly. Currently, good evaluators take a culturally responsive approach (Bowen & Tillman, 2015; Chon, Gero, & Treichel, 2015; Faulkner, Ayers, & Huq, 2015), which involves substantial efforts to ensure valid measurement with particular attention to cultural issues. Although a brief summary here cannot do justice to the complex process of culturally responsive evaluations, the main point is that evaluators must adapt their methods to fit the culture(s) of those involved in programs. Typically, this involves substantial collaboration with those who know the specific cultures and typically have life experience similar to those involved in the program. Among other things, the American Evaluation Association, in its public statement on cultural competence in evaluation (American Evaluation Association, 2011), has called for careful attention to these issues.

Use Measures That Can Detect Change

Sensitivity to change is a very important point in evaluations. Before a course, few students are expected to do well on a test covering the course material. After the course, however, students who mastered the material will do well. Classroom tests are good if they are sensitive to changes that occur as the skills of the students improve. Fairly stable characteristics of the students (such as intelligence) should not be the primary factors leading to differences among the students' classroom test scores. By contrast, the developers of intelligence tests seek to minimize the effects of specific learning experiences on intelligence scores and maximize the effect of general skills developed over a long time period. Evaluation of such stable traits is nearly opposite to the approach evaluators take when seeking to detect the impact of a specific program (Meier, 2004).

Furthermore, care must be taken to be sure that measures of outcome are affected by the program, not by a

great number of other influences. Community-level performance indicators were used in an evaluation of crime prevention efforts in Victoria, Australia (van den Eynde, Veno, & Hart, 2003). Unfortunately, the programs appeared totally ineffective because the outcome measures were not closely linked to program activities. For example, some programs (such as drama, wilderness experiences, training skills needed for circus performances) were designed to change youth culture. Even if effective, such an outcome would hardly affect emergency hospital visits a month or two after the program started; positive results might be seen over the course of a lifetime—obviously too long a time frame for any evaluation.

Several methods can be used to check on the expected sensitivity of a measure. Previous effective use of a measure with evaluations of similar programs would increase confidence that it is sufficiently sensitive. If the measure has been developed specifically for an evaluation, perhaps existing groups that should theoretically differ on the variable can be found. For example, a newly developed measure of scientific reasoning that reveals differences between physics and humanities majors is a better measure of an outcome of an innovative science curriculum than a measure that cannot discriminate between such groups (Lipsey, 1993). In estimating the sensitivity of a measure, evaluators must consider the typical score of the participants before a program begins. It is hard to detect improvements among program participants if most of them are already in a desirable condition when the program begins. This problem is called a ceiling effect. For example, in a given week very large proportions of hospital patients are discharged alive and most automobile drivers are accident-free; consequently, even when medical care improves or drivers become more careful, it is hard to detect changes in such variables unless observations are made on many people. If appropriate, one way to increase statistical sensitivity to small changes is to observe the same people both before and after participation in a program. Repeated measures designs increase statistical sensitivity and should be considered when appropriate (Lipsey, 1990; Shaughnessy, Zechmeister, & Zechmeister, 2008). Further, the best solution is to find other elements that show differences among people and predict the less frequent outcomes. For example, level of functioning and skill in driving would be more useful than only considering death and accidents.

Use Cost-Effective Measures

In planning program evaluations, the cost of developing and producing data collection instruments must be considered. Several principles are important to remember. First, efforts devoted to making test material attractive and easy to use reap higher response rates; computers make this job rather easy. Second, funds devoted to obtaining copyright materials are small compared to the costs of gathering, analyzing, and interpreting data. Third, because interviews are a much more expensive form of data collection compared to surveys, extensive interviewing should not be routinely planned. Some evaluation questions, however, cannot be answered using records or written surveys. At times, program participants are unable to use written forms (e.g., children, elderly, or very ill patients), need prompting to obtain complete information (poorly educated participants), do not respond to written surveys (physicians), or might be defensive when answering written surveys (addicts). Last, the relationship between the instruments and participant loss needs to be considered. An instrument that is in itself inexpensive may not be cost-effective if evaluators must spend a considerable amount of time following up tardy potential respondents or if so few respond that the results do not represent any group.

TYPES OF MEASURES OF EVALUATION CRITERIA

It is impossible to catalog the wide variety of measures of unmet needs or the ways to measure implementation and outcomes of programs. In this section several of the strengths and weaknesses of the most widely used techniques are described. Some ways to speed your access to the literature are illustrated.

Written Surveys and Interviews with Program Participants

One general point will come as no surprise to most readers—although the following sections address issues about a variety of ways to gather information from participants such as open-ended questions in contrast to forced-choice or Likert-type questions, one important choice when planning to collect information is whether to use hard copies or online methods. Although there are situations that still require surveys printed on paper,

increasingly, online options have substantial advantages that should not be dismissed quickly. Two obvious benefits of online surveys are the low cost of creating the surveys and making them available to participants and the reduced time and labor as well as the increased accuracy of data entry. As just reading about online options is not a full experience, see the eResources for some examples of online surveys and a list with links to some of the platforms that can be used for online data collection.

Written Surveys

Probably the single most widely used method of gathering data for evaluation is the written survey administered to program participants. The written survey provides the most information for the cost and effort required. Depending on the nature of the questions addressed, surveys differ in their reliability and validity. Asking clients of psychotherapy about the overall value of the service they received is not likely to yield particularly reliable or valid responses. However, the reliability of a survey that focuses on current specific behaviors is likely to be fairly high. Evaluators and stakeholders often draft surveys that are longer than they need to be and more complicated than participants are likely to complete. Gathering information that cannot be used is a poor use of evaluation resources and an imposition on the potential respondents.

Using surveys developed by others saves evaluators' time in writing and pretesting items. Furthermore, previous users of such surveys may have published the means and standard deviations that will provide comparisons to the program being evaluated, thus making the interpretation of findings more valid. When the program is indeed unique, evaluators are required to develop new surveys to measure attitudes about the program. Evaluators must be realistic about how much time they can afford to put into the construction of new instruments; like playing a guitar, writing survey items is easy to do badly.

Evaluators plan on reminding potential respondents to complete surveys. Reminder letters or telephone calls can be made soon after mailing a survey. Repeated reminders can be sent because with each reminder more surveys are returned. If too much time elapses before sending a reminder, many people who received the survey will have forgotten about it (Mangione, 1998). The proportion of people who will complete a survey varies with the effort required to complete it, the nature of the program, and the degree to which individuals identify with the program. Receiving less than 50% of the surveys back makes drawing most conclusions highly tentative in many cases. Even when participants respond at respectable rates, those who do not respond are probably different from those who do. In one study the 66% who returned a survey about their experience in counseling had participated in a mean of 20.2 sessions, but those who did not respond participated in a mean of only 9.5 sessions (Posavac & Hartung, 1977). If all had responded, the average of the sample would have been 16.6, not 20.2.

Interviews

Interviews are used when it seems that the members of the target population are unlikely to respond to a written survey, when respondents may not answer difficult or sensitive questions unless an interviewer is at hand to encourage them, or when evaluators are not at all sure what is most important to potential respondents (Denzin & Lincoln, 2005). The most compelling advantage that interviewing has over a written survey is the opportunity for the interviewer to follow up on the points made by the respondent. Interviewing is a difficult task requiring concentration and considerable familiarity with the interview questions and the program under study; it is not just chatting with program participants. Group interviews permit program participants or community residents to respond to the ideas of others; it is felt that such interaction yields richer and more well-thought-out ideas. Initial stages of needs assessments can often be profitably based on group interviews called focus groups (Krueger & Casey, 2000). Sometimes focus groups are used to gain information that can be used to design written surveys (Mitra, 1994). Interviews used in qualitative evaluations are less formal than is described here (see Chapter 8).

An alternative to face-to-face interviews is interviewing over the telephone (Lavrakas, 1998). Because telephones are a necessity of life in developed countries, an interviewer can contact just about anyone. Furthermore, telephone interviews do not involve travel, so their cost is lower than in-person interviews. However, the interviewer may have to make repeated calls and must establish rapport quickly before the

person who answers the phone classifies the caller as a salesperson and hangs up. A letter received before the telephone call explaining the need for the study and requesting cooperation may increase an interviewee's cooperation. Since both face-to-face and telephone interviews are more expensive than written surveys, choosing among them should not be done without careful consideration of cost and the nature of the evaluation being planned.

Checklists, Tests, and Records

Checklists

A list of various aspects of a planned program could be used to create a checklist to verify that implementation has been thorough. A list of desired behaviors or things learned by participants can also be used as a checklist to measure program outcomes. A complex observational checklist approach to evaluating the effectiveness of residential mental health treatment facilities was developed by Paul (1986). By carefully defining key behaviors, lists of specific behaviors (such as pacing) that are objective and visible were developed. Such checklists can be reliable and can be used very efficiently once observers have been trained and know the definitions of the target behaviors thoroughly. Greiner (1994) described checklists used in the evaluation of municipal programs such as street and building maintenance.

Individuals can be asked to use checklists to describe less objective behaviors in themselves or others. For example, family members might be asked to indicate the degree to which discharged rehabilitation patients perform various behaviors thus indicating how well people are able to take care of themselves. Mental health staff members might be asked to describe a patient's social functioning by checking one of several possible levels ranging from totally dependent on others to provide a supportive environment, through partially dependent (i.e., able to function with frequent therapeutic interventions), to totally independent (i.e., functioning well with no need to continue contact with a mental health center).

Tests

In many educational settings, measurements of cognitive achievement are frequently used to measure both needs and outcomes. Published achievement tests with high reliability are well developed and widely accepted in educational settings (Murphy & Davidshofer, 2005). As mentioned already, educational program evaluators should avoid choosing measures of aptitude or intelligence when looking for a measure of achievement. Evaluators in mental health settings can often find standard tests that reflect a patient's stress, anxiety, and depression, emotional states that should be relieved by most therapies. The rapid development of computerized databases has made the search for potential measures possible for an evaluator with access to a university library. A particularly useful resource for evaluators seeking measures related to medicine and mental health is the Health and Psychosocial Instruments (HAPI) database. Figure 4.2 contains illustrative citations obtained from HAPI. Each of the citations listed was just one of many that appeared to be relevant to the type of program listed. There are databases available relevant to many topics.

Agency Records

Detailed records are kept for employees of all businesses; medical care providers are required to keep incredibly detailed information on patients; in order to run manufacturing firms, orders, inventories, and completed products must be tracked accurately; and criminal justice agencies similarly keep detailed records. Many such records are in agency databases already. The outcomes of programs that are planned to improve health, decrease crime, raise productivity or quality, or increase skills among schoolchildren can often be traced in records kept by the organization sponsoring the evaluation. Because most information is kept for the purposes of working with individuals, summarizing information on a program level can provide insights not previously available.

Focus of the Program Being Evaluated	Name of Illustrative Measure	Example Citation Discussing or Using the Measure
Drug rehabilitation	Women in recovery Questionnaire	Turner, N. H., O'Dell, K. J., Weaver, G. D., Ramirez, G. Y., and Turner, G. 1998. Community's role in the promotion of recovery from addiction and prevention of relapse among women: An exploratory study. *Ethnicity & Disease, 8*, 26–35.
Juvenile delinquent rehabilitation	Multidimensional anger Inventory	Hemphill, J. F., and Howell, A. J. 2000. Adolescent offenders and stages of change. *Psychological Assessment, 12*, 371–381.
Depression among children	Children's depression Inventory	Silverman, W. K., Kurtines, W. M., Ginsburg, G. S., Weems, C. F., Lumpkin, P. W., and Carmichael, D. H. 1999. Treating anxiety disorders in children with group cognitive-behavioral therapy: A randomized clinical trial. *Journal of Consulting and Clinical Psychology, 67*, 995–1003.
Suicide prevention	Reasons for living inventory	Dean, P. J., Range, L. M., and Goggin, W. C. 1996. The escape theory of suicide in college students: Testing a model that includes perfectionism. *Suicide and Life-Threatening Behavior, 26*, 181–186.
Parenting enhancement	Responsibility for children Interview schedule	Leslie, L. A., Anderson, E. A., and Branson, M. P. 1991. Responsibility for children. *Journal of Family Issues, 12*, 197–210.
Reading for job skills	Ramsay corporation job Skills–office reading test	Impara, J. C., and Plake, B. S. eds. 1998. *The thirteenth mental measurements yearbook*. Lincoln: University of Nebraska Press.

Figure 4.2 Illustration of information on scales obtained from the Health and Psychological Instruments database. This database of scales greatly reduces the effort required to search for measurement tools.

PREPARING SPECIAL SURVEYS

Surveys are used so widely in educational, psychological, and policy settings that a brief overview of the principles of survey design would be helpful.

Format of a Survey

The format of a survey is an important factor in gaining the cooperation of respondents, analyzing the responses, and interpreting the findings. If a survey is to be self-administered, the layout must be attractive, uncluttered, and easy to use. For many evaluations, a structured answer approach format is preferable to using free response (or open-ended) questions because analysis of narrative answers is very time consuming. When the structured approach is adopted, it is still possible to produce a visually cluttered survey if the questions with different answer options are mixed together. For example, a mix of "Yes/No" questions, attitude statements with which respondents are to agree or disagree, and questions about frequency of past behaviors creates difficulties for respondents because they must repeatedly readjust to the response format as they move from item to item. The more difficulty a respondent has answering, the less likely he/she will complete a survey. Learning how to phrase a variety of questions in ways that permit them to be answered using the same answer format becomes easier with experience. The effort put into making the survey easy to use also helps evaluators when they seek to interpret and present the findings. The mean responses to items with the same answer format can be placed into a single table, thus making reports showing comparisons among items easier to interpret.

At times the topic of interest in an evaluation is complex and requires narrative answers. When the range of reactions to a program is not known or when having the respondents' own words might be of value, surveys with open-ended questions are often useful. Patton (1980) used both a structured answer format and open-ended questions when evaluating a performance appraisal system in an educational setting. Readers of Patton's report had access to both the statistical summary of the survey and the personal, often impassioned, free responses. McKillip, Moirs, and Cervenka (1992) showed that the most useful information for program improvement was obtained from items that elicited specific comments, such as "Name one thing you liked and one thing you did not like about [the program]." These open-ended, but directed, questions were more helpful than simply asking for comments. Evaluators planning an evaluation that makes extensive use of open-ended questions should consult materials on developing coding schemes for narrative material (e.g., Corbin & Strauss, 2008; Weitzman & Miles, 1995).

When preparing survey items, the most important principle is to remember who is expected to respond to the items. A statement that is clear to an evaluator might not be clear to someone reading it from a different perspective. Questions that have the best chance of being understood are written clearly, simply, and concisely. Such statements cannot be prepared in one sitting; they must be written, scrutinized, criticized, rewritten, pretested, and rewritten. Items written one week often do not seem acceptable the following week. Keep in mind that unless one is testing a skill, all respondents should be able to answer the questions. If the first draft cannot be improved, then the evaluation team has not learned to evaluate its own work.

There are several characteristics of clear survey items:

1 Good survey items are grammatically correct.
2 Items should avoid negatives because negatively worded sentences take more effort to read and are misunderstood more frequently than positively worded items. Double negatives are especially likely to be misunderstood.
3 Well-written items use short, common words.
4 Good survey items focus on one issue; an item such as "The training was informative and presented well" combines two issues. In this case it is possible to interpret a positive answer but not possible to interpret a negative answer. What should the director of training do if many people respond negatively —have a workshop on teaching techniques or one on the content of training?

Several practices can help detect survey items that need improving. First, the statements should be read aloud. Awkward phrasing is often easier to detect when read aloud than when reviewed silently. Second, imagine someone reading the item who dislikes you and may be looking for a reason to criticize your work. (No kidding, this will really help you to detect problems with any of your written material.) Third, have a colleague read the draft items. Fourth, think through how to interpret each possible answer to each item. As mentioned above, sometimes one answer can be interpreted but another cannot. Fifth, remaining ambiguities probably will be detected if the revised draft is administered as an interview with several people from the population to be sampled. Ask them to paraphrase each question; if they can, the question is clear.

Once survey items are prepared, they must be arranged into an order that is easy to use and appears logical to the respondent. The first questions should refer to interesting, but nonthreatening issues. The goal is to encourage the potential respondent to begin the survey. Questions dealing with controversial issues should be included later. If there is a concern that the order of items might affect the answers, consider making two different forms of the survey so that the possible effect of order can be examined. The demographic items (such as age, gender, or occupation) are best placed last since these are least likely to interest respondents. During an interview, however, demographic items should be addressed first to build rapport (Babbie, 2009). In all cases, the use to which the information is to be put should be explained.

Instructions and Pretests

The instructions accompanying a survey need to be clear since no interviewer is available to clear up difficulties. It is best to underestimate the skills of respondents and give more instructions and examples than might be necessary. Once the survey and instructions are prepared, it cannot be assumed that potential respondents will interpret instructions and items just as the evaluator intended. The best practice is to administer the survey to a small sample of people to learn how they understand the instructions and the items. After doing so, it is likely that the survey will need to be revised once more.

CASE STUDY 4

Primary Care Behavioral Health

Data Driven Solutions LLC (DDS) was retained by a local pediatric clinic to evaluate a pilot program integrating primary care services with behavioral health services. The pediatric clinic was opened in 2011 to serve children with exceptional health needs. Conceptualized as a "one-stop shop," the clinic houses medical, dental, hearing, vision, radiology, and behavioral health services, as well as a pharmacy. The parent agency has been serving vulnerable children for more than a century primarily through convalescent pediatric care for children who face significant medical, developmental, and physical disabilities, many of whom require ventilator support, and residential services for

abused, neglected, and abandoned children. The outpatient clinic is an expansion of the commitment to underserved children in the community. Case Study 1 describes the evaluation of a later expansion to this clinic.

The Primary Care Behavioral Health (PCBH) pilot was grant funded and designed to provide feedback in order to improve the program as the grant cycle progressed. Evaluators met with the primary investigator regularly to report on patient satisfaction and the prescribed outcome measures to ensure that the program was meeting the needs of patients and families in a satisfactory way and meeting targets for increasing services.

Meeting Needs

The evaluation noted that the National Institute of Mental Health reports that one in five children present a "severely debilitating mental health disorder," yet only 20% receive the mental health services they need. Further, the Kids Count Data Center estimates that 24% of Kentucky children aged 2 to 17 have one or more emotional, behavioral, or developmental conditions. Ensuring access to good-quality, integrated physical and behavioral health services for vulnerable children is ongoing and challenging. The PCBH pilot hoped to meet the needs of children by expanding behavioral health care access for children and their families, specifically by embedding a staff clinical psychologist 30 hours each week in the exam rooms suite of the clinic along with the primary care providers.

Implementation

Electronic medical records were audited for the year prior to the start date of the PCBH pilot and the number of mental health services provided before and after the pilot was implemented were compared, documenting the solid implementation. Additionally, the evaluators were able to assess how the data was entered, where it was stored in the record, and what difficulties there were with codes and other records of visits. The evaluators then met with staff to problem-solve procedural issues to address some of the identified challenges.

Stakeholders

The stakeholders who were included in data collection were child patients, their parents, and the primary care providers.

Side Effects

No negative side effects were detected. Positive side effects related to transportation issues and stigma. The intention of the program was, in part, to increase ease for families who might have difficulties with transportation to and from the clinic. This was only measured by administering the parent survey item regarding increasing accessibility to mental health at the same time as pediatric visits. It was also hoped that the introduction of behavioral health in a primary care setting would normalize behavioral health services and serve to reduce the stigma that often accompanies mental health issues. This was not measured; however, primary care providers gave feedback that patients and parents seemed to be comfortable with the combination of behavioral health in the primary care setting.

Improvement Focus

One particularly clear aspect of the evaluators' approach that illustrates an improvement focus was their detailed examination of the electronic medical records and their subsequent collaboration with staff to find workable solutions to address identified issues. The goal clearly was to help the agency learn from its experience with the system and move forward in a better way.

Outcomes

Mental health outcomes were measured for child patients, and parents completed a survey after each encounter with the behavioral health psychologist to determine their satisfaction with the mental health services provided. Primary care providers also completed a brief pre- and six-month post-survey assessing satisfaction with the PCBH pilot. In addition, a semistructured interview was conducted at the end of the pilot year for more in-depth qualitative information.

Results demonstrated that significantly more children were served with the psychologist embedded in the clinic than before the PCBH pilot program began. Qualitative data showed that the stakeholders had favorable impressions of the model and were hopeful that the program would continue. Although this case is not unique in reporting both quantitative and qualitative data, it is a good illustration of an evaluation that used multiple formats and sources of data, rather than relying on a single source such as patient self-report. Specifically, observational data on patients, self-report from parents, and surveys from providers gave a much more well-rounded summary than is typically reported.

Nuances

The support staff and primary care providers' attitudes about the PCBH were important but not measured. Perhaps the greatest unmeasured mediator is the rapport building and skills the psychologist was able to demonstrate with patients and providers, although data showing the psychologist was rated very highly by parents and other providers does suggest that the relationship may have been a crucial factor. Although many evaluations try to examine what factors may have been important mechanisms, it is quite common, as in this case, for the final report to acknowledge how potential explanations for the outcomes eluded measurement and analysis.

Summary

When single sources of data are the only basis for an evaluation, the possibility exists that the conclusions are the result of the choice of that source rather than the actual workings of the program. It is true that sometimes using multiple sources and methods can yield

conflicting and therefore confusing results. In this case, however, the data from multiple sources and formats generally confirmed each other, which supports greater confidence in all the findings.

Source: S. Cambron (personal communication, November 26, 2017).

EVALUATOR PROFILE 4

Stacy Cambron, PhD—Committed to Improvement

Stacy has been a mental health provider since earning a master's degree in medical social work in 1989. She went on to earn a doctorate in clinical psychology from the University of Louisville in 1998. She has experience working in a variety of settings across the country including private practice, community mental health centers, residential/forensic placements, and hospitals. Stacy also earned her Master of Public Health from the University of Louisville School of Public Health and Information Sciences in May 2015 with a concentration in biostatistics. The master's degree in public health is evidence of her philosophy of lifelong learning and staying current with cutting edge research methodologies.

In addition to consulting and private practice, Stacy recently joined the clinical faculty at Spalding University in the doctoral program in clinical psychology. She is active with the Kentucky Psychological Association and serves as a member of the Advocacy Committee. Having been a member of interdisciplinary teams throughout her career, Stacy is passionate about promoting a systems or integrated approach to behavioral and medical health care. Toward that aim, she has participated in workgroup meetings to become well versed in the current challenges that face entities dedicated to providing health care and social services.

Stacy is a proponent of utilizing evidence-based practices and outcome measures whether she is practicing psychology, teaching and supervising students, or evaluating the effectiveness of programs and interventions. Her recent focus on research design, evaluation, and statistical methodologies make her eager to assist organizations. Even the smallest project can yield great benefits if designed and implemented correctly.

Source: S. Cambron (personal communication, January 18, 2018).

SUMMARY AND PREVIEW

The use of measures of different variables from multiple sources of information is one of the core qualities of valid and useful evaluations. When choosing measures of variables, evaluators use the criteria for good measurement instruments to evaluate the specific approaches being considered. Is something important being measured? Is the approach sensitive to small changes? Does the measure seem valid, reliable, and cost-effective? Could the measure itself create changes in the variable being measured? In choosing variables and information sources, evaluators seek data sources that are not subject to the same limitations; instead, evaluators seek variables that provide information from different perspectives.

The selection of appropriate criteria for evaluation studies and the use of good ways to measure criteria are marks not only of competent evaluators but also of ethically sound work. As social scientists whose findings can have practical impacts on others, evaluators face a greater number of ethical dilemmas than those who study only theoretical questions. Chapter 5 deals with some of these ethical issues.

STUDY QUESTIONS

1. Consider a program setting with which you are familiar. You might think of the Financial Affairs Office of your college, a hospital laboratory, or a preschool center. Think of some survey items that the participants would be competent to answer and some that require specialized training to answer and thus should not be addressed to participants.
2. In that same setting identify the groups who might be surveyed or interviewed to obtain information on the quality of the program. On what aspects of the program is each group especially able to provide useful information?
3. Consider the variables and information sources that might reflect the quality of the program you have been considering. Arrange variables in two columns—place variables that would be expected to put the program in a favorable light in one column and variables that would be expected to highlight program problems in the second column. Compare the columns. How do those two sets differ?
4. Think of some evaluation issues that can only be answered using interviews and some that can be

answered best by survey items with set answer options (i.e., agree strongly through disagree strongly). How do these two sets of questions differ from each other?

5 Suppose that an unethical evaluator wanted to stack the deck to guarantee a favorable evaluation of a program. What types of variables would one use to achieve this? By contrast, suppose that someone wanted to guarantee an unfavorable evaluation. What types of variables would one choose to do that?

ADDITIONAL RESOURCE

Fowler, F. J. Jr., & Cosenza, C. (2009). Design and evaluation of survey questions. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods*, rev. ed. Thousand Oaks, CA: Sage.

This chapter provides a very readable description of the development of surveys with many illustrative items, good and bad. Many stakeholders who request evaluations seem to believe that anyone can write survey questions. The problem is that the findings from poorly prepared surveys are seldom interpretable.

5 Ethics in Program Evaluation

EXAMPLES OF THE PROBLEM

Evaluators often find themselves in ethical conflicts that are seldom experienced by social scientists engaged in basic research. Although the following scenarios are hypothetical, experienced evaluators can identify with the problems described below.

A Project That Cannot Be Done Well

Evelyn Marshall works for a social science firm, Evaluation, Inc. The firm's staff is interested in winning a contract to evaluate an early parole program recently approved by a state legislature. The central question to be answered is whether early parole leads to better outcomes (e.g., fewer rearrests) after release from prison relative to the outcomes of prisoners who serve a greater portion of their sentences. The wording of the request for proposals is clear: the legislature wants an evaluation that examines whether the program causes a lowered arrest rate after release. Further, the evaluation is needed within six months. Because early release is dependent on good behavior, Marshall knows that those who are released early are already viewed as less likely to be rearrested soon after release. A reasonable test of the outcome as defined should take at least several years to collect meaningful data. In this situation Marshall knows that she does not have enough time to test the effectiveness of the program. Should she and her colleagues prepare a proposal even though they know that it is impossible to do what the legislators want?

Advocacy versus Evaluation

Morris Franklin completed an evaluation of the Community Mental Health Center outreach program for high school students whose poor school performance is thought to be related to drug abuse. As typically occurs, Franklin found evidence that supports the program's effectiveness (all stakeholders like the program), evidence that does not support the program (grades of the participants did not go up), and some ambivalent evidence that could be interpreted as favorable or unfavorable depending on one's point of view (program students hold more part-time jobs than those not in the program). Franklin's report was received favorably by the center's director. Meetings were scheduled to consider the possibility of adding a tutoring phase to the program to help improve the participants' grades. Franklin felt that he had done a good job. Later that week, however, the director asked him to write a proposal to the state to permit extending the program to other high schools in the area. When Franklin mentioned the negative findings, the director told him to focus on the positive findings. "I want an upbeat proposal," he said. Would it be unethical to write the proposal but not mention the known weaknesses of the program that the evaluation detected?

STANDARDS FOR THE PRACTICE OF EVALUATION

The need for ethical standards in research has prompted a number of organizations to develop statements defining ethical principles to guide the work of research teams (e.g., American Educational Research Association, 2002; American Psychological Association, 2010; American Sociological Association, 1999). Because the findings of program evaluations may be applied soon after evaluations are completed, evaluators face many more situations calling for ethical choices than do basic researchers, and many of these choices are different from those faced by laboratory scientists. Consequently, several statements of principles have been

prepared specifically for the practice of program evaluation (Joint Committee on Standards for Educational Evaluation, 1994). Figure 5.1 includes the ethical principles adopted by the American Evaluation Association (2004).

1. Systematic Inquiry: Evaluators conduct systematic, data-based inquires.
2. Competence: Evaluators provide competent performance to stakeholders.
3. Integrity/Honesty: Evaluators display honesty and integrity in their own behavior, and attempt to ensure the honesty and integrity of the entire evaluation process.
4. Respect for People: Evaluators respect the security, dignity, and self-worth of the respondents program participants, clients, and other evaluation stakeholders.
5. Responsibilities for General and Public Welfare: Evaluators articulate and take into account the diversity of general and public interests and values that may be related to the evaluation.
Note: Each principle was followed by several subpoints. The statement was ratified by the membership of the American Evaluation Association, July 2004.
Source: www.eval.org/d/do/594. Retrieved January 31, 2018.

Figure 5.1 Ethical principles adopted by the American Evaluation Association.

This chapter includes material based on the statements of ethical conduct in research, as well as descriptions of good program evaluation practices. Ethics in evaluation means more than respect for research participants and honesty with money and data. Evaluators have the responsibility to provide clear, useful, and accurate evaluation information to the stakeholders with whom they work. Furthermore, evaluators seek to work in ways that have the potential to improve services to people. Working in settings designed to help people means that reports might have immediate relevance to the organization sponsoring the program whereas the work of social scientists will go through a period of refinement before being utilized by organizations; for that reason, errors in basic research are less likely to harm people immediately. By contrast, poorly done evaluations have quickly affected the provision of services to people, disrupted the staffs of service organizations, and encouraged the use of harmful, novel medical treatments. Ethical concerns, thus, relate to all stages of an evaluation—from initial planning through the presentation of the results to interested parties. In the hypothetical example above, Morris Franklin learned that there may be crucial ethical challenges even after the evaluation has been completed.

As you will see in the following sections, ethical issues are divided into five categories: treating people ethically, recognizing role conflicts, serving the needs of possible users of the evaluation, using valid methods, and avoiding negative side effects of the evaluation itself.

ETHICAL ISSUES INVOLVED IN THE TREATMENT OF PEOPLE

The first responsibility of an evaluator, as it is with the basic researcher, is to protect people from harm. Since harm can be done to people in a variety of ways, concerned evaluators guard against harm to all people associated with a program. Further, federal regulations specify detailed rules and procedures to follow when conducting research with people. These regulations are known as 45 CFR 46, which refers to the Code of Federal Regulations, Title 45, Part 46 (U.S. Department of Health and Human Services, n.d.).

Detecting Ineffective Programs

Often the first issue evaluators face concerns whether any harm can come to someone receiving the program that is being evaluated. Although medical, educational, and social service programs are offered with the purpose of helping those who participate, sometimes programs have either no impact or a negative impact. Few social service and education programs have as dramatic negative side effects as some medications have

had (Mosher & Akins, 2007). Yet, attractive ideas appear that turn out to fail to help and indeed hurt those who hoped to benefit. Most people agree that self-esteem is related to doing well, but the ways to help children develop self-esteem are not well understood. It now seems that evaluations of teaching methods centered on increasing children's self-esteem have shown that merely increasing self-esteem does not lead to increased achievement (Bronson & Merryman, 2009). The students like teachers who encourage them, but do not correct their homework, but less is learned compared to a more structured teaching approach. These findings might partially explain the continued lag in mathematics and science achievement of American children compared to children of other nations.

Obtaining Informed Consent

Another way to protect people is to obtain prior agreement from program participants who are to take part in an evaluation. This is especially important when evaluators plan an evaluation that includes random assignment of program participants to different forms of treatment or to a control group not receiving the treatment. When people are asked for their agreement, it is important that their consent be informed—that is, they must understand the request before giving consent. Informed consent means that potential participants themselves make the decision about whether to participate, and that sufficient information about the program be provided to enable them to weigh all alternatives. If a person is misled or not given enough information about the risks involved, then informed consent has not been obtained, even if a person has signed an agreement to participate.

Attempting to provide enough information to enable people to give informed consent can create an additional ethical dilemma for an evaluator. Revealing too much information about a new form of service can create expectations on the part of the participants or demoralize those not selected for the new program (Cook & Campbell, 1979). When informed consent procedures have the potential to change the behaviors of the people in the evaluation, then the validity of the evaluation can be threatened. There is no clear way to resolve such a conflict. One approach, included in the American Psychological Association's ethical principles, is to consider the potential harm to the participants. If the potential for harm is low and the harm, if any, is very small, the need for the evaluation may make it ethical to conduct the evaluation. Sometimes, those not getting a service immediately can participate in the program at a later time.

Maintaining Confidentiality

Information gathered during an evaluation is to be treated with the utmost care so that the privacy of program participants or managers not be violated. The confidentiality of information can be protected in a number of ways. Identifying data with a person's name is not always necessary. If matching information from various sources is necessary, evaluators can use identification codes that only a particular respondent would recognize, such as the first name and birth date of the respondent's mother. If contacting the respondent later is necessary, the project director alone should keep a master list of the respondents' names and addresses and the code identifying the respondent. Some evaluators working with very sensitive information have stored the names and codes in a different country. Few evaluators deal with information that sensitive; however, confidentiality once promised must be preserved.

ROLE CONFLICTS FACING EVALUATORS

Evaluators gather information for the purpose of assessing the quality of program plans, the fidelity of implementation, or the value of outcomes. Since people serving on the staff of programs earn their living from their work, it should not be surprising that conflicts can occur between evaluators and program staff. The clients served by a program also have a stake in the conclusions of an evaluation; if the program is changed, they may lose a needed service. The stakeholders of publicly funded programs include government agencies and taxpayers. Although few taxpayers are aware of evaluations or even, for that matter, the programs themselves, taxpayers' interests are served when programs achieve objectives efficiently. In an important sense, many ethical issues arise from conflicts of interest among the stakeholders involved with the program. Ideally, evaluators serve all stakeholders even though evaluators are usually employed by only one stakeholder—most

often the organization management. As mentioned in Chapter 2, forming clear ideas of who wants the evaluation conducted and how all the stakeholders could be affected by an evaluation is crucial. The stakeholders can be expected to have different, even conflicting, interests; anticipating these conflicts can spell the difference between a poorly received evaluation due to unresolved disagreements among stakeholders and a carefully balanced, albeit controversial, evaluation.

Identifying the stakeholders of a program is an important ethical task. Imagine an evaluator was given the task of conducting an evaluation of a new method of sentencing convicted criminals. Suppose that it was widely believed that the sentencing practices of judges were too lenient and varied widely from judge to judge. Suppose also that the state legislature took the determination of length of jail sentences for certain crimes out of the hands of judges: a defendant judged guilty of one of these crimes now faced a sentence determined by law, not by the judge. The desired outcome of such a law would be a lowered crime rate because those inclined to crime would know that if caught, the sentence will not be light. Who would be the stakeholders for an evaluation of such a law?

First, the state legislators are stakeholders because they wrote the law that was being evaluated. Those state legislators who voted for the law likewise would have wanted a favorable finding. By contrast, some legislators who opposed the law may have preferred to see their position vindicated by a finding that the law did not have the intended effect. The governor would have an interest in the outcome of the study because she or he may have proposed the law and would want all evaluations of the work to be favorable in order to enhance chances of reelection. All citizens who look to the courts to help in preventing dangerous individuals from committing additional crimes would be stakeholders as well. Judges, too, would be affected because their day-to-day behavior was being studied. Prison officials needed to be aware of the evaluation findings, especially with respect to the possible influence of the new sentencing procedures on their facilities. Police officers also would have been interested to learn how the new sentencing law had been implemented. Defense lawyers, defendants, and the state attorney's office would also have stakes in the outcome of the evaluation because their strategies might depend on how sentences were determined. Clearly, many groups care about the findings of program evaluations.

Evaluators should try to minimize potential conflicts among the stakeholders before the evaluation begins. Evaluators are subject to less pressure if they can negotiate agreements on issues such as who will have access to the findings, what information will be used, and how different patterns of findings will be interpreted. If the study begins without settling these questions, the probability that different groups can manipulate the evaluation to suit their own purposes increases. At times, suspicions and disputes about ownership of information have made it impossible to conduct a credible evaluation. Individuals who do not want an evaluation conducted benefit from such a result; all others, however, lose.

Public conflicts among stakeholders complicate the work of evaluators; moreover, less visible conflicts also exist. Some writers argue from a social justice ethical point of view (see Rawls, 2000) that evaluators should examine the program assumptions and outcomes in order to learn if justice is served by the program and how it is administered. Evaluators try to detect when programs lead to undesirable outcomes for some stakeholders that might be overlooked. Guiding Principle E, Point 5 (see Figure 5.1) includes this point: "[E]valuators will usually have to go beyond analysis of particular stakeholder interests and consider the welfare of society as a whole."

RECOGNIZING THE INTERESTS OF DIFFERENT STAKEHOLDERS

Considering the interests of different stakeholders is important for making an evaluation as useful as possible for all those who may use or be influenced by an evaluation. Figure 5.2 presents an older yet still relevant example that illustrates how stakeholders' views can lead to different conclusions, even when interpreting the same information. Focusing on the needs of one stakeholder group can easily lead to narrow and misleading conclusions (McGarrell & Sabath, 1994).

Program Managers Are Concerned with Efficiency

Those who are responsible for managing an organization are concerned about the efficiency of its operations. For-profit firms must produce goods and services in an efficient manner; otherwise, competing firms will win

a greater proportion of the market. In time, inefficient firms go out of business or are bought by more efficient ones. Nonprofit human service agencies do not operate on the same principles as private firms; however, effective managers seek to provide the most service possible within limited agency budgets. In a manager's judgment, the most important aspect of an evaluation may frequently be information on the efficiency of the program. Chapter 12 covers cost-effectiveness questions important to managers.

Was Workfare a Success or a Boondoggle? It Depended on How You Looked at It.

Workfare is an approach to providing assistance to unemployed people. Instead of welfare benefits, the person is given a part-time, temporary job with pay coming from the government. This mechanism was developed in the US in the late 1960s and continues in many countries. While people look for work, they spend a week or two a month in public or nonprofit organizations doing jobs that would not have been done by regular employees.

The program remains controversial. Some critics said that the program was short-sighted, that the people did not learn real job skills, and that the government did not adequately supervise these temporary employees. Other critics said that workfare merely required welfare applicants to "sing for their supper," making welfare programs punitive. A public official countered by saying that some workfare participants needed workfare to learn that holding a job requires coming to work on time, staying all day, not getting into fights, and bathing before coming to work. Nonprofit service organizations reported that the day-care, janitorial, and food service assistance they got through workfare had improved their services to other poor people.

An evaluator would have been hard pressed to gather data that would have been acceptable to all stakeholders, because there are basic values conflicts that data could not have resolved.

Figure 5.2 An illustration of how stakeholders' values can conflict. (From Brotman, B., "Workfare": What state terms success others call boondoggle. *Chicago Tribune,* Sec. 3, pp. 1, 4, January 2, 1983.)

Staff Members Seek Assistance in Service Delivery

The needs of the staff are best served if the evaluation can provide practical guidance, improving the effectiveness with which they serve clients, students, patients, or customers. An evaluation of the flow of information through a large division of a firm may reveal inefficiencies. If the staff is provided with viable alternatives to the current procedures, they may be well served by an evaluation. An evaluation of teaching effectiveness in a college can help identify the areas in which new faculty members have the most difficulty and, thus, lead department chairs and college deans to develop preventive actions to assist new faculty members before serious problems fester (Angelo & Cross, 1993). An evaluation of a medical residency program can estimate not only the proportions of unnecessary medical tests ordered by medical residents, but the reasons for such errors in judgment. In this way education programs can be improved and residents can serve their future patients more effectively (Posavac, 1995).

Evaluators also know that an evaluation is a good vehicle for recognizing the good work of staff members. Evaluations that focus on shortcomings to the exclusion of providing positive feedback are seldom as valued as more balanced ones.

Clients Want Effective and Appropriate Services

The people for whom a service is designed frequently have no voice in the planning and implementation of either programs or evaluations (Bowen & Tillman, 2015; Morris, 2015; Sturges, 2015). This oversight means that the group for which the program was developed is often the least consulted. However, participants are not unified, they seldom have spokespersons, and they do not hire evaluators. Among the ways evaluators can fulfill their responsibilities to participants is to compare the participants' needs with the service offered, to help the staff and manager better understand those needs, and to structure recommendations around both needs and strengths. The interests of the manager and staff have sometimes taken precedence over those of the

participants. It should be noted that there is no reason to assume that program participants—students, patients, drug abusers, or trainees—know fully what they need, but that does not mean that their views can be ignored.

Another aspect of the stake participants have in an evaluation is their interest in continuing to receive service during an evaluation. Conducting an evaluation without causing any disruption in normal service is usually impossible; however, ethical concerns require that service disruption be minimized.

Community Members Want Cost-Effective Programs

Most service agencies receive some of their financial support from local community residents through taxes and contributions, or indirectly through reductions in property taxes often enjoyed by nonprofit organizations. In some ways the local community is in a position similar to that of the people served: The community is dispersed and does not hire evaluators. Human service agencies often receive further support from national or state bodies, charitable organizations, and foundations. These groups certainly have a stake in the success of the programs they support. Frequently government offices and foundations commission evaluations of programs receiving their support. In these instances their interests are probably well protected. Ethically planned internal evaluations also reflect the interests of the groups that supply financial support for the service agencies.

THE VALIDITY OF EVALUATIONS

After potential harm to participants is minimized, possible role conflicts explored, and stakeholder needs identified, evaluators can turn to ethical issues associated with the validity of an evaluation project. Conducting evaluations that are not appropriate to the purposes for which they were commissioned is just as unethical as, for example, not protecting the confidentiality of information obtained from participants. The following sections cover the ethical ramifications of four of the most frequently found threats to the validity of evaluations.

Valid Measurement Instruments

Evaluators in educational and mental health settings frequently use standardized, published tests when measuring the expected outcomes of programs being evaluated. The most well-developed tests are standardized achievement tests designed to estimate a child's progress in school. Because these tests are so well developed, there is a temptation to use them even when they may not be appropriate to measure the outcomes of a program. In other words, one cannot simply ask if a measurement tool is valid because validity varies depending on the setting and the specific program participants. Choosing an inappropriate way to measure a hypothesized outcome can obscure the effects of a program (Lipsey, 1990; Lipsey, Crosse, Dunkle, Pollard, & Stobart, 1985) or, even worse, lead to a misleading conclusion. For example, in an evaluation of a course on ecology, a standard achievement subtest on science that included questions on hygiene, biology, and earth science was used as an outcome variable (Joint Committee on Standards for Educational Evaluation, 1994). This subtest was not appropriate for measuring the achievement of the students in the innovative ecology course. The evaluation's negative findings did not reflect the actual outcomes of the program. The school district would have been better off had the evaluation not been conducted at all; the program was saddled with a negative evaluation that—although not valid—had to be explained by program supporters.

Skilled Data Collectors

It is likely that even people with little training can competently administer a standardized, pen and paper test; however, many evaluations use information collected through personal interviews. Interviewing is not an easy job. Good interviewers possess interpersonal skills and common sense that permit them to obtain the information needed while maintaining the goodwill of the person interviewed. It requires a degree of maturity and a respect for truth to be able to record and report attitudes at variance with one's own.

Skill in gaining the cooperation of people is also required of good interviewers. Interviewers need to

exercise polite persistence as they seek to speak with potential interviewees. Carey (1974) interviewed terminally ill patients as part of an evaluation of hospital care for dying patients. It is difficult to conduct such interviews without appearing to be more concerned about getting information than the welfare of the patients and their families. Children present a very different challenge. Interviewing children in a compensatory education program designed to improve self-confidence, study habits, and social skills would require patience and skill to gain the cooperation of the children. An interviewer without experience in working with children may learn very little about their feelings and attitudes because children do not respond to social expectations as adults do. Indeed, if young children do not feel at ease, they might not say anything at all to an interviewer!

Appropriate Research Design

One theme of this book is that the research design must fit the needs of those who might utilize the information sought. As later chapters illustrate, different evaluations require more or less scientific rigor and more or less deep understanding of the participants. However, once an information need has been articulated, conducting the evaluation would be unethical if it is known at the outset that the planned evaluation cannot answer the questions that the sponsors asked. One would hope that this would be unusual; however, Vojtecky and Schmitz (1986) concluded that few evaluations conducted on health and safety training programs provided accurate information on program outcomes. This seems related to the dilemma faced by Evelyn Marshall, as outlined at the beginning of this chapter; it is unlikely that she can conduct an evaluation that answers the questions the state legislature had initially specified.

At this point Marshall has at least three alternatives: (1) go ahead with the project and use her skills to make it seem as though the completed evaluation answered the questions originally posed, (2) decline to conduct the evaluation and hope that she can find an alternative contract to keep herself productively employed, or (3) negotiate about the actual questions to be addressed by the evaluation. The first alternative is unethical and the second is risky for Marshall; however, the third may well prove productive. Frequently sponsors of evaluations are unskilled in research methods and do not know when they have given an evaluator an impossible assignment. It may be that the legislature does not really need an evaluation that relates changes in parole procedures to later parolee behaviors. What is needed may be only a careful documentation of the actual implementation of the new law and stakeholders' reaction to it. If negotiations reveal that implementation is the primary issue, Marshall can conduct the evaluation; in fact, she can do a better job than she would have been able to if she had to devote effort trying to convince others that she was answering the initial questions.

Adequate Descriptions of Program and Procedures

A basic characteristic of science is its public nature. Science is to be conducted so that others can evaluate the procedures used and repeat the research. In this way, mistakes and misinterpretations can be detected. Frequently evaluations are not described in sufficient detail to permit others to understand either the program or the evaluation procedures. Patton (1980) illustrated the advantage of describing the implementation of a program as part of a program evaluation. He found that the evaluators of a program to help young low-income mothers learn parenting and financial management skills concluded that the program was ineffective when, in fact, the program had not been implemented. If the evaluators had observed and described implementation, they would have learned that the planned program had not been put in place. A large proportion of evaluations have not addressed implementation issues (Lipsey et al., 1985; Shadish, 2002; Summerfelt, 2003). It would be very valuable to show that degree of faithful implementation of a program was related to better outcomes (Saunders, Ward, Felton, Dowda, & Pate, 2006).

Besides describing the program, the evaluation procedures should be presented in enough detail so that others can understand how the evaluator obtained and analyzed information. As will be mentioned later, not all interested parties want to know all the details of the evaluation procedure. However, such detailed reports should be available to people who are considering implementing the program elsewhere or who want to compare evaluations of similar programs from different settings. If it is impossible to compare evaluations because they are inadequately reported, one might wonder if the report writers were clear in their own minds about what was done.

AVOIDING POSSIBLE NEGATIVE SIDE EFFECTS OF EVALUATION PROCEDURES

The ethical issues in this section are less central to the conduct of basic research; however, in a program evaluation these questions can take on immense significance both for evaluators and for stakeholders affected by the evaluation study.

Can Someone Be Hurt by Inaccurate Findings?

Inaccurate findings can show either falsely positive findings—erroneously suggesting that the program is effective—or falsely negative findings—erroneously suggesting that the program is not effective. When such false conclusions are due to random statistical variation, the first false finding is called a *Type I error*, and the second is termed a *Type II error*. There is clear evidence medical research sometimes provides support for treatments that are later found to be of marginal or no value (Ioannidis, 2005); the positive findings of the original evaluations may have been due to Type I errors.

Such false conclusions can also be made when insufficient attention is paid to the design of an evaluation. Without careful thought, evaluators can focus on the wrong variables or use too short a time span to show either positive or negative effects. At other times an evaluator's enthusiasm for a program may lead to falsely optimistic conclusions. Inadvertently encouraging the use of a harmful program due to insufficient care in conducting an evaluation is an ethical problem with which basic researchers need rarely be concerned; however, it is often an important issue for evaluators.

In contrast to falsely favorable evaluations, falsely negative evaluations can also harm people by encouraging the elimination of beneficial services. An overlooked aspect of a summer preschool program for children from low-income families, involving mothers in the education of their children, led to a negative evaluation (Lazar, 1981). Showing mothers how to work with their children and introducing them to a school system encouraged them to develop a level of involvement with the education of their children often not regularly found among lower income families. Evaluations that focused only on the intellectual growth of the children during the eight-week program were too narrow. The falsely negative conclusions could have harmed other children if other programs had been eliminated.

Consider Statistical Type II Errors

Type II errors, being unable to conclude statistically that a program is effective when it is, in fact, effective, can occur because the sample of program participants involved in the evaluation was small or just happened to be atypical, and when the measures of outcome were of low reliability. Evaluators observe samples of participants, seldom whole populations, and the information sources available to evaluators are never perfectly reliable. Basic researchers worry about Type II errors because conducting research that yields inaccurate conclusions is a waste of time. However, Type II errors in a basic research study typically cause no harm to others. Evaluators, on the other hand, have an additional worry about random statistical errors: they do not want to conclude falsely that a valued program is ineffective. Thus, evaluators are much more concerned about Type II errors than are basic researchers (Lipsey et al., 1985; Sackett & Mullen, 1993; Schneider & Darcy, 1984).

Unfortunately, evaluators work in situations that make Type II errors very likely. In an attempt to reduce the demands made on program participants providing information, short, thus less reliable, surveys might be used and the number of variables might be limited. To reduce the disruption of services, only a few participants may be tested. A study of the effects of oat bran on cholesterol levels involved only 20 healthy participants (Swain, Rouse, Curley, & Sacks, 1990). It should not have been surprising that no effect was found; follow-up evaluations showed that the conclusion of the initial evaluation was a Type II error (see Ripsin et al., 1992). After reviewing 122 published program evaluations, Lipsey et al. (1985) concluded that the research designs of a large proportion of evaluations were too weak to detect even a moderately sized effect, not to mention a small effect. Two ways to reduce Type II errors are to use large samples and outcome measures with high reliability (Lipsey, 1990).

Pay Attention to Unplanned Negative Effects

Ethical evaluators are careful to examine the programs as implemented, not just as designed. One aspect of this issue is the low level of actual implementation of programs, as discussed earlier. A second aspect of this issue involves possible negative side effects of a program. Just as physicians are concerned about side effects of medications, evaluators work most effectively when they are alert to unexpected negative effects of a program. For example, using food stamps may embarrass recipients, prison regulations may cause dependency leading to more difficulty in making decisions after release, and an arbitrary method of introducing safer working conditions may alienate employees. Such outcomes, of course, are not found in the goals of the programs as planned.

Because program planners and managers might not anticipate the possibility of negative outcomes, evaluators can make useful contributions if they detect important negative side effects of programs. A sexual harassment educational program given without prior announcement at regularly scheduled university department meetings had an adverse effect on male participants who felt they were being accused of doing wrong (Bingham & Scherer, 2001). Recognizing a negative side effect can lead to an improved program or manner of presenting one. Figure 5.3 includes some additional negative side effects from policies with desirable goals. Sieber (1981) and Howard (1994) describe at length negative side effects from policies and laws that added to the very problems that they were supposed to alleviate and created new problems. What is most critical is that the evaluator not hide negative side effects.

DESIRABLE GOAL	PROGRAM	NEGATIVE SIDE EFFECT
Reduce auto pollution	Require pollution control devices	Reduced gas mileage
Save energy	Burn waste sawmill material for heat	Loss of nutrients that could have been returned to the soil
Reduce the use of nonrenewable energy sources	Process corn into ethanol to use in American automobiles	Food prices increased thus reducing the food supply for third world populations
Improve food production	Use fertilizers and farm machinery	Poor farmers in poor nations who could not afford the cost of machinery sold farms and became unemployed city dwellers
Raise quality of public school teachers	Test public school teacher competence	Insulted good teachers; implied that meeting the minimum standard defined good teaching

Figure 5.3 Serious negative side effects unexpected by the program developers.

Analyze Implicit Values Held by the Evaluator

Conflicts between the role of the program advocate and the role of evaluator have already been mentioned. Other unexamined values may be hidden in statistical analyses. It is seldom mentioned that a common statistical procedure such as presenting an average (e.g., the mean) improvement requires the implicit assumption that the benefit for one participant is as valuable as the benefit to other participants. An early evaluation of the impact of watching "Sesame Street" examined the overall mean achievement of children and concluded that the series was successful in achieving some of its goals (Ball & Bogartz, 1970). However, when the child viewers were divided on the basis of socioeconomic background, a positive correlation was discovered between improvement in cognitive skills and income level: the gap between the children of lower and upper income groups actually increased (Cook et al., 1975). Although "Sesame Street" was designed particularly to appeal to children of families with lower incomes, it appeared more effective with the children of more well-

to-do families. In other words, even though the program had a positive impact, it did not provide the special help for disadvantaged children that was a fundamental goal of the program. Without a creative reanalysis of the evaluation, this important—but subtle—finding would have been overlooked. Some writers argue that when program benefits are uneven, the less privileged group should benefit more (Bunda, 1983). What is important for evaluators to recognize is that merely examining overall effects implies endorsement of a simplistic interpretation of the utilitarian ethics of Jeremy Bentham, who argued that the most ethical outcome would lead to the "greatest happiness for the greatest number" (Waller, 2005). Even Bentham recognized disparities among people (Atkinson, 2009). Regardless of what choices are made, it is always best to deal with ethical issues in an explicit manner.

INSTITUTIONAL REVIEW BOARDS AND PROGRAM EVALUATION

Students beginning a project for a program evaluation class frequently ask whether they need approval from their university's Institutional Review Board (IRB) before beginning. IRBs are required by federal law to ensure that medical and behavioral research studies are conducted ethically and research participants are protected from harm. Colleges, universities, hospitals, and other agencies sponsoring research are required to have a committee that examines research plans to be sure that participants are treated in compliance with ethical standards. Do evaluators need to seek the approval of IRBs? There is disagreement among evaluators about this. Newman and Brown (1996) "strongly advise" evaluators to obtain IRB approval. Surely no harm will be done to an organization if IRB approval is sought. However, requiring IRB approval for program evaluation would lead to rather odd limitations on the activities of program managers.

Imagine an arts and sciences college that began special, low-enrollment sections for first-year students. Suppose that these sections, called "freshman seminars," were supposed to (a) encourage interaction among class members and (b) orient students to college life while still covering the essentials in the existing courses (i.e., psychology, history, political science, etc.). The college curriculum committee, the dean, and the admissions office would be interested in knowing whether students noticed a difference between freshman seminars and other courses and whether they found the experience valuable. Does the dean need IRB approval to evaluate the implementation of and the reaction to freshman seminars? It would seem strange to say that the dean cannot evaluate part of the curriculum for which he/she is responsible. In fact, 45 CFR 46 explicitly recognizes educational organizations' rights to evaluate their own programs (U.S. Department of Health and Human Services, n.d.). Furthermore, it would seem strange to say that the dean could evaluate these courses, but that someone working in the dean's stead would need approval by a committee.

Suppose a student in a program evaluation course has the opportunity to complete a class project for a manager of a human resources department in a nearby company. It would seem odd for the student to need IRB approval in order to complete a project that the manager wants done and, in fact, supervises. IRB regulations were not designed to control these situations. IRBs are not charged with supervising how managers evaluate activities in their own agencies. Instead, IRBs are charged with supervising *research*, especially research done with publication in mind. Many IRBs have concluded that evaluations such as those just described do not fall under their purview and, thus, those IRBs would not seek to examine proposals to carry out such evaluations. Reports of these evaluations are not going to be presented beyond the organizations sponsoring the activities being evaluated. Not requiring IRB approval does not mean that one can be reckless or careless about how program evaluations are carried out. The dean and the human resources manager keep student and employee records confidential as part of their normal work. All people working under the direction of such managers are expected to follow such ethical practices as well.

The situation changes if a faculty member or college student is *not* acting as the program manager's agent. Suppose that a student approaches a program director requesting to carry out a program evaluation for a master's thesis. Although the project may be a program evaluation, the project looks a lot more like research instead of an aspect of responsible program management. In such a case, IRB approval is needed prior to gathering data.

Last, exercise common sense: evaluating a sex education curriculum for a middle school is much more sensitive than evaluating a new scheduling procedure for college students. Evaluators of programs in socially controversial areas (such as drug rehabilitation or sex education) would be wise to consult with someone

familiar with evaluations in such areas even if there is no IRB in the agency sponsoring the program.

Ethical Problems Evaluators Report

Morris (2015) has recently updated earlier findings regarding the ethical challenges that evaluators say they have faced themselves. Although readers are encouraged to examine the detailed table Morris presented of 25 distinct problems, two overarching themes are clear: the stakeholders requesting the evaluation exert pressure on the evaluators to misrepresent the findings in one of a fair number of ways, or participants contributing to the evaluation are not treated properly. The first problem clearly subverts the entire evaluation process, so it functions as one of the most destructive difficulties evaluators face. The specifics may vary: some stakeholders communicate their expectation that the evaluation will reach specific conclusions, whereas others attempt to prevent evaluators from learning certain information such as evidence of negative outcomes. But whether the interference is explicit or implicit, good evaluators are prepared either to persist with an accurate approach or to decline to conduct such evaluations.

Related to the second theme of participants being mistreated, a number of writers have noted that such practices clearly follow patterns of power, where stakeholders with little economic or other forms of power are often the ones mistreated (Morris, 2015; Sturges, 2015). At times the mistreatment may appear more subtle, such as when those who receive the program being evaluated do not have a clear option to provide useful data, whether they are not asked or are assessed by inappropriate measures (Bowen & Tillman, 2015). Good evaluators find appropriate ways to respond to such challenges to protect people and the integrity of the evaluations. And the good news is that, most commonly, ethical challenges in evaluations are of modest proportions, and actions well within most evaluators' skill-sets can often lead to productive resolutions, as in the following case study.

CASE STUDY 5

Bridges to Tomorrow

In 2007, Metro United Way of Louisville, Kentucky, supported by multiple donors, began a major initiative called the Bridges to Tomorrow project to address child readiness for school with a broader approach than most projects with similar goals. Recognizing that children's ability to succeed in educational settings is affected not only by individual factors such as motivation but also by important environmental factors such as family stability, the project targeted children's experiences in child care settings by addressing staff training and their family's economic situation by providing coaching that was especially prepared to address financial issues. (A later version of the financial coaching component was the subject of Case Study 14.) The Bridges to Tomorrow project was implemented in four community centers in Louisville that provided similar but not identical programming. For example, three of the sites were community centers with existing child care facilities. A fourth center was included despite never offering child care partly because it provided a greater range of racial and other diversity such as clients speaking at least six languages other than English.

Program evaluators Dr. DeDe Wohlfarth and Dr. Ken Linfield at Spalding University began to work with the Bridges project in 2008, providing annual reports from 2009 to 2013. The project involved a number of groups of contributors, some of whom worked independently and had no contact with some others. In particular, a team from another, nearby university conducted the regular formal assessments of the children enrolled in the child care centers that were a part of the Bridges project and presented their own annual reports. Early in the project, the assessment team and the evaluators drew slightly different conclusions from the same data regarding one of nearly 20 analyses—the evaluators used a slightly relaxed level of probability for determining statistical significance due to the limited statistical power and the desire to avoid Type II errors (saying there is no effect when, in reality, an effect is present) in addition to Type I errors. A subsequent report by the assessment team indirectly took issue with the approach used by the evaluators.

Although the program staff did not perceive a major conflict, in an attempt to address the situation in a helpful way, Dr. Linfield contacted one of the members of the assessment team and met, starting with more casual introductions and then proceeding to talk about the project and the teams' approaches. As a result, the teams first came to an agreement on a common set of parameters for analyses, so their reports in subsequent years were completely consistent. In later years, although the two teams continued to submit their separate reports, they also collaborated on a joint Executive Summary. Although this is an example of a relatively minor resolution of conflict, it illustrates that direct conversations that find and build on common ground are often much more productive than other, more contentious, approaches.

Meeting Needs

The evaluation reports noted the nested needs—a broad need for higher quality child care, many low-income people need better knowledge and skills with regard to financial matters, and a coaching approach is one of the best ways to engage people for long-term change. Addressing all three of these needs at the same time is more likely to build children's readiness for school in a sustainable way. As virtually all stakeholders in the project already accepted the reality and scope of these needs, the main points were made without extensive documentation of details.

Implementation

Earlier reports provided timely descriptions of the elements that had been fully implemented as planned, as well as the minor changes in some details and the common experience of recruitment of participants occurring at a somewhat slower pace than hoped. Later reports summarized earlier points and provided the perspective of a longer time frame, such as noting that the modifications could be best understood "as a refinement, not a departure, from the original program goals" (Linfield, Wohlfarth, Gibson, & Jones, 2013, p. 2).

Stakeholders

Although most reports noted the various groups of stakeholders implicitly, the final report included data from "Interviews with Key Stakeholders," noting that these "included Metro United Way Staff (e.g., project manager, CEO, etc.), Vendors/Contractors (e.g., teacher mentors, program evaluators, etc.), and Agency Partners (e.g., community center program directors, coaches, etc.)." (Linfield et al., 2013, p. 12, emphasis in the original). Further, representatives from the agencies contributing support were present at regular program board meetings throughout the years and especially at the annual meeting.

Side Effects

Comments on side effects were included in some, but not all, reports, noting, for example, that enthusiasm and mutual support among participants at the same centers was an unexpected positive benefit, and that the emphasis on data collection was substantially distracting to some participants.

Improvement Focus

The reports and the frequent interactions between the evaluation team and program staff and administrators were consistently framed toward improvement, noting, for example, the understandable basis for problems with data collection in a service-oriented project. As with many programs with multiple reports over several years, the earlier reports were explicitly formative while later reports focused more on summative results.

Outcomes

The reports addressed a wide range of outcomes, from the formal assessments of children's readiness for school in six areas, to details of the quality of the child-care centers, to parents' knowledge and attitudes about finances, to the nature of goals parents determined and the extent of progress toward those goals. As is common, results were mixed with clear evidence of change in some of the areas, but no evidence for improvement in others, although no findings showed deterioration. For example, the child-care centers showed nearly universal, substantial improvement in quality to the point of reaching a ceiling, whereas, despite important work toward financial and other goals, parents showed minimal change in their financial status.

Nuances/Mechanisms

One of the evaluators' expectations was that an important mechanism in this project would be participants' progression in their readiness to change according to the trans-theoretical model of stages of change (Prochaska & Norcross, 2001), so they added measures to assess participants' baseline and later readiness. Qualitative data from focus groups and other sources appeared to indicate substantial movement from pre-contemplation to action. But one important challenge in the project was that helping participants with their immediate situation took precedence over data collection, so missing data reduced the statistical power available. Thus, those participants for whom data was available did not show clear evidence of progression on the measures as a group, and the changes documented were not significantly associated with the positive outcomes.

Summary

This multilayered, complex program evaluation demonstrates several key takeaway lessons. First, programs with multiple goals and intentions need multiple assessment measures. Although simplicity is generally ideal, multiprong programs do not typically allow for such simplicity. Second, a key component of success in program evaluation is working collaboratively to resolve differences of opinion, scope, or data analysis. Being direct without being overbearing can help navigate even challenging interpersonal relationships. Finally, complex program evaluations rarely lend themselves to short sound bites of information. Although the goal of program evaluation is to present data in clear, simple, understandable language to key stakeholders, sometimes the answer to whether the program is meeting its goals is, "Yes, sort of. In some respects, yes, and in others, no. Mostly it depends."

Sources: Prochaska, J. O., & Norcross, J. C., *Psychotherapy,* 38, 443–448, 2001; Linfield, Wohlfarth, Gibson, & Jones, 2013.

EVALUATOR PROFILE 5

DeDe Wohlfarth, PsyD—Being a Bridge

Dr. DeDe Wohlfarth is a licensed clinical psychologist, specializing in children, and the Director of the Child/Adolescent/Family

Emphasis area in the PsyD program at Spalding University, where she is a Professor in the School of Professional Psychology. She explains that, for her, being a program evaluator is not another role that she assumes, but is simply another way to describe what she does, just like "professor" and "clinical psychologist." Whatever the label, she sees herself as a bridge—a bridge between the university and the community or a bridge between research and people, ordinary people.

Dr. Wohlfarth sadly acknowledges that there is currently a chasm between researchers and everyday people, partly because too many researchers express themselves in ways that have to be translated before they connect with regular people. She notes that empirical evidence has the potential to make enormously valuable contributions to people's lives, because the right choices need to be based on "meaningful data," not just gut feelings, which can easily be wrong. For example, she recently led a workshop on parenting and based many of the specific points on clear principles of behaviorism that have been identified through rigorous research. Many people, however, raise their children based on "gut feelings" and ignore research that might be helpful to them.

Dr. Wohlfarth notes "numbers are not bad"; instead, they are tools to help us. We have to ensure, however, that numbers make sense and are relevant for people. DeDe understands this frustrating situation, noting that she is the parent of four teenagers, so anything she uses has to work in the real world. One particular area needing better translation is when findings are based on one particular population and need to be applied to many diverse populations. Addressing these limitations requires knowing the limits of otherwise good tools and understanding how to adapt research when the data fall short. Working at Spalding with its focus on social justice has both helped her refine some of these techniques as well as provided a particularly good setting to use them. As she notes, Spalding is located in one of the ten poorest zip codes in America, so "we don't have the luxury of being an ivory tower."

Interestingly, like Ben Birkby (see Profile 1), DeDe had a Graduate Assistantship with Bob Illback at REACH when she was in the PsyD program at Spalding, and her dissertation was a program evaluation. Although it had not been the most obvious choice for her, given her passionate interest in clinical work with children, she once had asked Bob, "What do I need to do to be part of the future?" Bob's advice was to develop "a diverse set of skills," and she did. She offers one explanation of why such proactive choices are not more common with a favorite quote: "The problem is not that we can't read the writing on the wall. The problem is we think it is addressed to someone else."

Source: D. Wohlfarth (personal communication, January 25, 2018).

SUMMARY AND PREVIEW

The major ethical issues involved in conducting evaluation studies fall under five topics: the protection of people studied, the danger of role conflicts, meeting the diverse needs of different stakeholder groups, avoiding threats to the validity of the evaluation, and staying alert to negative side effects that may be related to the program or the evaluation itself. The applied and political nature of the settings in which program evaluations are conducted create more ethical dilemmas for program evaluators than are experienced by basic researchers. Although these conflicts and limitations require evaluators to exercise considerable care in carrying out evaluations, these issues also make the role of the evaluator vital and exciting.

This chapter completes the general, introductory section of this text. The next topics concern the specific types of evaluations. Measuring needs for programs and developing methods to monitor programs are essential tasks in themselves, but they also help clarify the conceptual foundations of programs. These aspects of program evaluations are addressed in the next two chapters.

STUDY QUESTIONS

1. Analyze the different views of stakeholders who would be involved with the development of a teacher evaluation procedure in a college or university. Be sure that you do not assume that all students have the same interests or that all faculty members would react to such evaluations in the same ways.
2. Contrast the problems that Type I and Type II errors create in basic research versus those in program evaluation, with special regard to ethical considerations.
3. Some evaluators do not mention all negative or unfavorable findings in official evaluation reports, but do bring such issues to the attention of program managers during private meetings. What are the reasons favoring such a practice? What problems might this practice create or encourage?
4. Suppose you have been asked to develop a report on the wishes of employees of a nonunionized company concerning benefits. After a day of interviewing you bump into an employee whom you do not know as you are leaving the plant. He tells you that the company's general manager has told the employees to hide their true feelings and thus mislead you into thinking that they are quite happy with the way things are. Because employees do not want to rock the boat, he says that everyone is going along with that idea. The employee hurriedly walks off. If he is telling the truth, what ethical problems does that give you? What should you do next?
5. Imagine how to deal with pressure to violate confidentiality.

a It is not unknown (although not common) for evaluators to be asked to identify individuals who made certain statements. Consider how to keep this problem from ever coming up.

b Suppose someone alleges that someone else in the organization is engaged in criminal behavior. It could be misappropriation of funds. What ethical duties would you have to the accused, to the accuser, and to the organization sponsoring the evaluation and the program?

ADDITIONAL RESOURCE

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

A variety of program evaluation scenarios that call for ethical reflection are presented. Many leaders in program evaluation comment on the situations and describe what they would do in those situations. Often we learn best when we see what experienced people do as they apply principles in specific situations.

6 The Assessment of Need

Human services and educational programs are developed to serve people in need and to facilitate positive development. Congress appropriates funds for federal nutrition agencies, small towns sponsor volunteer fire departments, and school districts build and staff schools because communities have decided that needs of people will be unmet without such agencies and services. How do we decide that there is a need to be met through community action? Who decides? What information can be gathered to demonstrate the level of community need? Perhaps the *wants* of a population must be changed so that it recognizes its own *needs* (Weiss, 2002). How does an evaluator work with planners who select programs to be offered?

Perceptions of need are greatly influenced by local and national media. A series on local crime presented on TV news caused people to feel less safe even if they had experienced no crime themselves (Heath & Petraitis, 1986). Descriptions of flu led some parents to take their children with colds to emergency rooms. Groups that attract the attention of media often create the sense that their concerns are more important than those of others (Morgan, Fischhoff, Bostrom, Lave, & Atman, 1992; Slovic, 1993). Such reactions may well mobilize opinions, but judgments about what services are needed should be based on quantified, representative information.

People mean many different things when they say that they need something. In this chapter, a definition of need is developed to remind evaluators to check on how the term is being used. Next, the most widely used methods of studying needs for human services without ignoring a community's strengths are reviewed. In settings in which an internal evaluator is likely to be called upon to evaluate a program under development, internal evaluators might seek to be part of the planning committee. Involving evaluators in the development effort provides planners with an independent perspective as they formulate program goals and impact models.

DEFINITIONS OF NEED

Many analyses of need do not include a definition of need. When evaluators and program staff talk about assessments of need, they usually refer to measuring a discrepancy between what is and what should be. Roth (1990) pointed out that there are at least five discrepancies that people could have in mind when they speak of needs. There might be a discrepancy between the actual state and (a) an ideal, (b) a norm, (c) a minimum, (d) a desired state, or (e) an expected state. In social service settings, it is probably safe to assume that discrepancies from ideals are seldom being discussed. But sometimes discrepancies from a norm are used to define a need as occurs when schoolchildren of normal aptitude read below their grade levels. This approach, however, leads one to conclude that people do not have needs when they enjoy an actual state that equals or exceeds the norm, minimum, or expected states (Scriven & Roth, 1990). This would lead to the odd conclusion that people drinking orange juice each morning do not need vitamin C because they have enough; a discrepancy definition, if used alone, is not helpful. Most assuredly, people do need vitamin C because without any we would, in time, become ill. Although needs assessments are not always explicitly linked to program theory, knowing the theoretical reason for an intervention can guide the needs assessment.

It is likely that many evaluators have been able to conduct needs assessments without a clear definition of need because program planners possess an implicit understanding of what is meant; nevertheless, it seems wise to adopt an explicit definition. Scriven and Roth's definition has merit: need refers to something (X) that people must have to be in a satisfactory state. Without X they would be in an unsatisfactory state; with X they achieve but do not exceed a satisfactory state. For example, people need access to a health care worker when they accidentally cut themselves severely enough to require stitches. McKillip (1998) describes a difficulty

with this definition. While we can all agree that someone needs stitches, we draw different conclusions about needs as we move from community to community. Most people in the world live in situations in which a bachelor's degree is not needed. However, in nations with knowledge-focused economies a college education is very important and, indeed, needed to hold a large proportion of jobs. Further, such needs may also change over time.

A teenager may want a cell phone that can transmit images, music, and text to gain status and convenience, but not having one would hardly make his life unsatisfactory. However, a real-estate executive may need a cell phone for client services because without it, she may lose enough business to place her in an unsatisfactory economic state. In other words, a need cannot be defined without regard to the social context in which a person or group lives and works. Consequently, McKillip (1998) concluded that the decision that an unmet need merits a programmatic response is the result of "… *a value judgment* [not an objective assessment] that some *group* has a *problem* that can be *solved*" (McKillip, 1998, p. 263). Note also that only problems that we think we can solve are treated as unmet needs, which is a good match when evaluating the program that is intended to meet those needs.

Lurking beneath some of the confusion about the definition of need is the distinction between things people need on a regular basis (such as oxygen or vitamin C) and things needed only once (such as a high school diploma). Another confusion occurs between those community resources whose use is fairly predictable (such as elementary schools) and those whose use may be predictable for a whole community, but whose use cannot be predicted on an individual basis (such as a trauma center). Individuals may believe that they will never need a trauma center (and most people will be correct); however, the community at large may need a trauma center. Those intrigued by various approaches to needs should study the table of categories and specific meanings of need in Watkins and Kavale (2014).

When estimating the need for any type of program, distinguishing between the *incidence* and the *prevalence* of a problem is helpful. Incidence refers to the number of people experiencing a problem during some time period, such as a year; prevalence refers to the number of people who have the problem at a given time. For example, the incidence of the common cold is high: most adults contract at least one cold each year; children contract more. However, most people recover within 10 days, so the prevalence at any one time is low. The distinction is important: a response to a problem will be different depending on whether a problem temporarily affects many, or is a long-lasting problem for a few people. For example, attempts to help unemployed people will differ depending on whether it is believed that the unemployed are mostly between jobs or are likely to be out of work for a long time.

Note that the definition of need does not rely on people knowing that they have a particular need. People may be unaware of a need (an iron deficiency), deny a need (alcoholics usually deny a need for rehabilitation), or misidentify a need (some adolescents desire drugs when they need social skills). Of course, most people know when they need assistance with housing, education, health care, or employment. Unfortunately, there is no single procedure or one source of data that provides all the information required to estimate needs. Still, one important pattern is that those served by programs are less frequently involved in defining needs than more influential stakeholders (Altschuld & Watkins, 2014; Watkins & Kavale, 2014).

SOURCES OF INFORMATION FOR THE ASSESSMENT OF NEED

The major sources of need-relevant information include objective data about the health, educational, social, and economic conditions of a community, facts and opinions that the residents themselves provide, and the conclusions of experts who know the community well. Before heading out with surveys in hand, evaluators should address several preliminary issues. First, who are the people whose unmet needs are being studied? Second, what are the resources currently available to these people?

Describing the Current Situation

Population to Be Studied

The people to be studied should be identified before gathering any data. The population might be the residents of a state, the pupils of a school district, or the employees of a factory or business. Instead of defining

the population as all members of a geographical area or organization, evaluators might narrow their definition to unemployed adults, homebound elderly, unmarried pregnant teenagers, or high school dropouts. Some public agencies may be limited in whom they may serve. Private organizations such as hospitals and businesses may have fewer limitations and thus more difficulty in identifying the people they wish to serve with a new clinic or new product. A university opening a branch campus would want to know how many of its current students live in the area of the proposed campus, how many area students will graduate from high school during the next decade, and how many people in the area are currently enrolled in continuing education programs. Some information describing the population would be easy and inexpensive to obtain; sometimes we can just ask people what they need. A management support company might use a self-report survey with its own employees containing items such as the illustrative activities in Figure 6.1. Job skills employees say that they need and want training in would provide one view of unmet needs for job training. Such information would be supplemented by other views of training needs.

The needs assessment report might show that part of the organization or community has great unmet needs, whereas other subgroups require no additional attention. Learning how the people seek to meet unmet needs would be important. Employees of the hypothetical management services company in Figure 6.1 might seek help from fellow workers causing inefficiency. In a community setting, some people might spend a great deal of time seeking assistance from agencies not designed to provide the needed help.

An evaluator can contribute to the planning process by being concerned with putting numbers on unmet needs. It is one thing to talk about the need for services to juveniles in trouble with the law and quite another to know the number and percentage of community male adolescents aged 16–19 arrested each year. Saying "The community's elderly residents lack easy access to social opportunities" is less informative than reporting "2,200 residents over age 75 have no access to public transportation, and half of them would require transportation in order to participate in services offered to them if the services were housed in the town hall."

Current Resources in the Community

One common problem in needs assessment is ignoring the strengths of communities and residents when developing programs (Altschuld & Watkins, 2014; Fine & Vanderslice, 1992). Indeed, the concept of needs assessment has been criticized for a myopic focus on problems without regard for how existing resources could be strengthened (McKnight, 1995). Social program planners might take a hint from businesses: it would be unwise to locate a new store or restaurant in a community without noting the location of established businesses. Thus, a careful cataloging of current programs, employers, health care facilities, churches, schools, and community organizations should be included in a needs assessment. The cooperation of such organizations can make programs more successful in reaching people in need, and the integration of existing services and organizations can lead to more comprehensive opportunities for eligible residents.

Examine the activities listed below. Indicate how much you use each one in your job with Management Support, Inc. Then indicate the activities in which you think you need more training.	How **often** is this activity needed in your job? (Check appropriate column.)			Do you want **more training** in this activity? (Check appropriate column.)		
	Never	Some-times	Often	No	Maybe	Yes
Developing budgets						
Monitoring budgets						
Managing records (Excel)						
Selecting new employees						
Orientating and training employees						
Conducting performance appraisals						
Selecting services to offer to clients						
Making & developing client relationships						
Developing proposals for clients						
Working directly with clients						

Figure 6.1 A self-report of training needs could be part of an assessment of the need for employee training in a company.

Social Indicators of Need

Communities needing additional social services probably display many dysfunctional characteristics. Trying to measure the degree of dysfunction—or its converse, well-being—is difficult. One approach is to track variables, such as crime and school dropout rates, that are symptoms of social problems (Felson, 1993; MacRae, 2009). Just as symptoms of physical illness can lead physicians to diagnose an illness, social indicators can suggest the underlying social problems that should be addressed (Carley, 1981; Ferriss, 1988). When a social problem is alleviated (as occurred with the advent of Medicare), social indicators should show a lower degree of social distress and dysfunction.

More and more, information can be found online to address at least some aspects of most needs. U.S. census data, crime statistics, information on health and disease, and more can be located using Internet searches. The eResources provide both descriptions of major categories of information and active hyperlinks to some ways to access that data. It is also important to note, however, that local information, when contrasted with regional and national data, can provide a powerful argument that there are local problems that might be solved or at least relieved by new or refined programs.

It is important to remember that indicators of problems do not reveal how those problems can be solved. Planners need a theory connecting the indicator to an unmet need as well as one connecting a possible intervention to the hoped-for improvement. With physical health, a high temperature indicates a problem. However, the problem causing the temperature is not obvious. Without a valid theory connecting elevated temperatures with a specific illness, ineffective remedies may be used. Recall that during American Colonial times, a flushed face and high temperature were thought to indicate the presence of too much blood, an invalid theory that led physicians to bleed sick people.

Using social indicators to identify community problems and to shape programs requires knowing the relationship between the indirect measures of the problem and the underlying social, educational, or economic

malady. Gaining such knowledge is certainly not easy. What malady underlies the decline in achievement in high schools? Some say TV watching and video games, entertainments that divert attention away from reading; others, the absence or inattention of parents who fail to take an interest in school work; and still others, unimaginative teachers. The alleged maladies suggest different remedies. Even if we agreed on the malady, sometimes several different responses are still suggested. If many teachers are not teaching well, perhaps teacher salaries should be raised to attract more competent people, perhaps we need to test teachers, or perhaps teacher certification requirements are keeping out people with wide experiences who lack the traditional credentials, but may become excellent teachers.

Do limitations on the use of social indicators mean that they cannot be used? No, but the limitations do mean that social indicators cannot be used as the only source of information about the needs of a community. Social indicators are like physical symptoms; they can tell us that problems exist and where they are most pressing. They cannot tell us what the fundamental causes of problems are or what society should do about them. Other sources of information can assist evaluators and planners to be more specific about the causes of the underlying problems and the options for alleviating them.

It is important to note that social indicators can be corrupted. Fraudulent social indicators are ultimately discredited, but often not until policies have been affected. Information is increasingly difficult to control as the development of Internet resources have multiplied; however, as with all data, evaluators do well if they carefully assess the credibility of social indicators before using them. With the wide variety of information available, it may become harder to purposefully distort social indicators in developed, knowledge-intensive countries than in the recent past when access to information was more easily controlled. Unfortunately, false reports spread rapidly as well. This warning has become increasingly relevant as the charge of "fake news" is used more and more to try to discredit reports that one simply dislikes.

Community Surveys of Need

A straightforward way to estimate community needs is simply to ask people about their needs. The residents of a community have opinions about the development of human services and about the particular services needed. Their opinions can be part of the information used in planning. If a service is relevant to all residents, the community should be surveyed systematically to obtain a representative sample of the opinions of the residents. Obtaining a truly representative sample is extremely difficult and thus expensive. A possible compromise is to use intact groups that are likely to be fairly representative of the community. Such a strategy would not be as expensive because intact groups completing a survey do not require much time from a survey administrator. Public schools and church groups are organizations that attract participants from wide sectors of a community. Depending on the nature of the program being planned, special groups of likely users could be sought. Thus, an interviewer gathering ideas about the needs of the elderly might use more detailed questions when a household including elderly people is found.

A survey to assess needs is often used to estimate the magnitude of need and the acceptability of particular approaches to meeting needs. Preparing a survey requires much care. Because completed surveys look simple, few people realize how much effort is required to write clearly, avoid leading questions, and obtain the information desired (Schwarz & Oyserman, 2001). When assessing needs it is easy to write questions that encourage respondents to endorse plans; one could list various services and ask if the services "should be available." If we were asked whether our hometown should have agencies able to provide counseling for residents having a personal or family crisis at any time of the day or night, we might answer "yes."

Although such questions appear to get right to the heart of the matter, they are inadequate in at least three ways. First, the questions do not deal with unmet needs, but with possible solutions to problems. As mentioned in Chapter 2, planners often fall into the trap of dealing with potential solutions before fully understanding the unmet need being addressed. It would be best to include questions on residents' opinions about the extent of community problems and about their own problems. Second, asking whether a service "should be available" would be a leading question. Attitude measurement research has shown that people tend to agree more readily than disagree with questions (Babbie, 2009); this bias is called *acquiescence*. To avoid writing leading questions, it is helpful to use questions that require responses other than agreement or disagreement. Asking people for examples of problems or to compare the relative extent of two different

problems might be useful approaches. Third, the direct question suggested above fails to mention costs of providing the service or to acknowledge that providing one service means that another cannot be provided. When respondents are asked what they would pay to support such services or additional safeguards, they are less enthusiastic (see Viscusi, 1992); however, research has shown that answers to questions about willingness to pay for services may be misleading (S. S. Posavac, 1998).

Services Already Available

People with unmet needs often seek assistance from a variety of sources that provide services related to what they believe they need, or they travel far to obtain services, or they do without the assistance that could benefit them. Evaluators can contrast the extent of need estimated with the level of services currently available in a community. Furthermore, a thorough analysis of the services available assures planners that new services would not be duplicating the work of existing agencies.

A planning group exploring the need for additional mental health services should identify all agencies, public and private, that provide at least some therapy or support for people with mental health or substance abuse problems. Thus, clinics, hospitals, courts, Alcoholics Anonymous chapters, school drug-abuse counselors, public health visiting nurses, church-based counseling centers, primary care physicians, psychiatrists, and social workers should be considered in the effort to learn about the people being treated in some way for emotional problems.

Planners may be required to work directly with agency files in order to develop a demographic description of clients. When this is necessary, absolute guarantees of confidentiality must be made and kept. When contacting such agencies for estimates of caseloads, it would be helpful to ask about additional providers of care. Developing estimates of the number of people using services is particularly difficult when services are provided informally, as child care often is.

Information concerning people needing a service, but not getting it, would be very hard to obtain since few in community agencies would be aware of such people. One way to estimate the numbers of such people is to examine agency reports of individuals who sought assistance but did not receive it because they did not meet qualification requirements, could not afford the fees, or had problems that did not fit into the services provided by the agency. Love (1986) illustrated another way to include unserved people in a needs assessment. Love used case studies of troubled teenagers to show how the structure of the welfare and mental health services in metropolitan Toronto made it impossible for an adolescent to get help even though his alcoholic father had thrown him out of his home. The unstable and, at times, aggressive 17-year-old had several recent addresses in different communities and needed several different services. It was unclear where he should have sought help, and only a few agencies could assist him with his multiple problems. Although several compelling cases do not show the extent of unfilled needs, case studies illustrate vividly that service reorganization or expansion may be needed to relieve certain types of needs.

Key Informants

Key informants are people who know a community well and could be expected to know what needs are going unmet. Finding such people is not easy, however. One way to start is to meet with professionals whose work brings them into contact with people who may have unmet needs to be served by the program being considered. In addition, people active in the community, such as clergy, politicians, or YMCA leaders, can be approached for information. All key informants can be asked for recommendations of people who might also have information. A sample formed using such recommendations is called a "snowball sample" (Babbie, 2009) because new participants are added to existing ones somewhat like when rolling a large snowball. Guidance counselors in schools might have some good ideas for needed changes in vocational school curricula, and they might be able to recommend human resource directors in local businesses who would have valuable information on strengths and weaknesses of recent high school graduates. For health issues, physicians, clergy, clinic managers, school nurses, and social workers would all be important. Although surveys can be mailed to such people, a far better level of cooperation can be obtained if they are interviewed personally.

Key informants are sought because they have the closest contact with residents in need; however, this close contact is also the source of their most glaring weakness. Since people in need are memorable and since key

informants may see a great number of such people daily, the numbers of people in need are often overestimated. This tendency is called the "availability" bias because the person making the estimate can easily recall individuals who fit the category (Heath & Heath, 2008). Consequently, psychologists are likely to overestimate the number of people needing counseling and remedial reading teachers the number of people needing literacy tutoring.

A related problem with key informants is that their expertise may lead them to view a problem in ways that are quite different from how others view the problem. Give a child a hammer, and the child begins to act as though everything needs to be hammered upon. Teachers view education as the answer to social problems, lawyers view legal services as a primary unmet need, and trucking company executives are not likely to view railroads as a solution to shipping needs. Having expertise in a particular form of service makes it likely that informants would readily think of ways of meeting unmet needs that are related to their expertise. A premature focus on any one alternative leads to overestimating the likelihood that the alternative is the correct one to relieve unmet needs (see Posavac, Sanbonmatsu, & Ho, 2002). This is a reason why planners should consider several alternatives before choosing a course of action.

Focus Groups and Open Forums

In addition to the option of administering needs assessment surveys to representative samples of community residents, two approaches can be used to encourage residents to share their opinions with others.

Focus Groups

Marketing research firms use small informal groups to discuss new products and ideas for new products (Malhotra, 2010). Although the kinds of questions addressed to focus groups are similar to the open-ended questions used in interviews, a discussion among 7–10 group members prompts participants to reflect more thoughtfully about the reasons for their opinions than they would if interviewed alone. The focus group leader tries to remain in the background to encourage a relaxed atmosphere and to foster sharing of opinions among members; however, the leaders must be sure participants keep on the topic. Focus groups usually consist of people who are similar to each other in age, income level, and education. Since the goal is to foster a free exchange of opinions, great differences among members could lead some members to guard their comments. Although similar to each other, the members of a focus group should not know each other prior to the meeting and, most likely, will not see each other afterward. Some incentive for attendance is usually given since invited participants must travel to the meeting location and spend, perhaps, two hours in the discussions.

Focus groups can be used in any phase of an evaluation; however, they serve particularly well in learning about the reactions to potential services or changes in current services of community agencies or private organizations. Krueger and Casey (2000) described using focus groups to discover why agricultural extension classes were drawing fewer and fewer enrollments among farmers in Minnesota even though a written survey suggested that farmers were interested in such classes. It could have been assumed that financial troubles were responsible, but focus groups told a different story. Farmers said that they were interested and that they could afford the fees; however, they wanted to be sure that the instructors were knowledgeable and that the material was applicable to their farms. In addition, it was learned that personal invitations would be more effective than the printed flyers that had been the standard way of announcing the classes. After changes had been implemented, attendance increased markedly.

The questions addressed to focus groups tend to be open-ended questions. Figure 6.2 includes some questions that could have been used in focus groups of rural business people. Plans were being made to develop an assistance program for small rural businesses as part of a state-funded project to encourage rural economic development. Extension classes for owners and managers of rural businesses were being considered, but before such classes could be developed, colleges needed to identify the information needs of business owners. These questions were designed to reveal those needs. See Case Study 8 for an example of the use of focus groups and some of the areas of questions used.

1. Over the past few weeks, did you have an experience in your business that caused you to seek information? Describe that experience and tell us where you went for the information.
2. Over the past year, think of times in which you needed help in managing your business; jot down those situations.
3. We have a list of topics suggested by some other owners and managers for programs to help them in their businesses. Look over this list and compare it to yours. Please report which items are the most important to you. (When a topic is suggested, the leader can ask the next question.)
4. What makes the area of [topic] important to you? (After such a question is discussed, others will have suggestions that will also be discussed.)
5. A topic that was mentioned by several people was [topic]. Where would you go to learn something about it?
6. People get information in different forms; how would you prefer to get the information? (The leader is to probe to clarify answers such as: in-person contact, phone calls, newsletters, meetings, workshops, classes.)
7. What makes a provider of business information credible? How do you know when you can trust the information you get?

Figure 6.2 Focus group questions to learn about the information needs of owners of rural businesses. Note that discussion among the members occurs after each question. (Adapted from Krueger, 1994.)

Failing to seek the reactions to planned programs may lead to offering services that are not wanted. An English department faculty member developed a series of workshops for non-English faculty members in her college to show how they could help students improve their writing. She obtained funds from the dean, but failed to assess whether faculty would attend the workshops. At the first workshop there were more graduate student program evaluators present than faculty participants. If a focus group of faculty members had been presented with the plans, she might have learned that few faculty members were interested in changing how they taught their classes or in attending these workshops.

Open Forums

An open forum is another method based on group interaction. Unlike focus groups that are selected on the basis of some common characteristics, participants of an open forum are self-selected. Often a governmental agency will announce the date of a community meeting to consider some planning issue. In many cases there are legal requirements to have such meetings. Those who learn of the meeting and choose to attend can participate. Sometimes participants' comments are limited to perhaps three minutes.

The advantages of open forums include unrestricted access for anyone who wishes to participate, the very low cost, and the possibility that good ideas may be offered. One major disadvantage is that self-selected groups are not representative of a community because many people cannot or will not attend such meetings. Once when a town held an open meeting to discuss how to spend a small federal grant, only one resident attended; he urged the development of a children's park along a railroad right-of-way, a location the planners viewed as dangerous. A less obvious disadvantage is the possibility that public discussions may raise expectations that something can and will be done to meet the needs discussed. It is also possible that assertive individuals with personal grievances can turn the meetings into gripe sessions producing little information on the needs of a community (Mitra, 1994).

Users of focus groups or open forums need to be cautious in drawing conclusions from the views of specific individuals. The needs of individuals can be quite compelling to casual viewers of TV news programs; in fact, even people trained as researchers have been known to put aside the findings of carefully conducted research and adopt conclusions based on the personal, compelling experiences of one or two articulate individuals rather than more valid, but abstract, information (Rook, 1987).

INADEQUATE ASSESSMENT OF NEED

The reason to assess needs is to improve the quality of program planning. Without understanding needs, programs cannot be designed appropriately nor can program objectives be developed wisely. When objectives are not specified clearly, it is impossible to learn if the program achieved what was wanted. When needs and the context of needs are not assessed accurately, programs and services cannot be as efficient or effective as possible. Furthermore, providing programs designed to meet needs that are already being met is not a good way to spend resources. The most common problems are as follows: sometimes needs are not measured, the context is not understood, the type of implementation required is not considered, and the denial of need is not considered. In contrast, a thorough assessment of the most obvious needs often uncovers other, sometimes more important needs (Lepicki & Boggs, 2014; Wedman, 2014).

Failing to Examine Need

It might seem a truism to say that program designers should be sure that proposed programs are planned on the basis of the needs of the people to whom the services are directed. It should be unnecessary to make this point, but sometimes programs are implemented without careful attention to unmet needs. Administrators at a college were concerned about the high attrition rate of freshmen who withdrew from college even though they were in good academic standing. As tentative ideas about possible program changes were being considered, it became apparent that the focus was on the desire to retain students, not on meeting the needs of the students who withdrew. Administrators had only speculated about the students' unmet needs and how program changes might prompt more students to reenroll for the sophomore year. Perhaps the attrition rate was being caused by personal and family situations that the college could not influence, or perhaps the students developed interests in areas of study in which the college did not offer courses. Attempts to lower the attrition rate among such students were bound to fail because no retention program could meet such unmet needs.

Another example of failure to examine need occurred in 1993 when it was proposed that the U.S. federal government provide free immunization vaccines for all children under 18. This was proposed because it was believed that cost was a barrier to immunization. However, when poor children—the least immunized group—are covered by Medicaid, immunizations are free already. According to the GAO (Chan, 1994, 1995), the reasons some poor children go without immunizations include parental disinterest and physician lack of initiative in immunizing children when the children saw their physicians for other needs. The planned, and nearly implemented, program would not have led to higher immunization rates because the reasons why some children did not obtain immunizations had been assumed, not identified empirically.

Failing to Examine the Context of Need

It is quite possible to have an accurate understanding of a community's need, but to fail to assess the community's capacity to support the program or the cultural context in which a program would be implemented (McKnight, 1995). One of the tragedies of the relationship between the developed nations and the undeveloped nations during the last several decades is the repeated failure of foreign aid initiatives to reflect the economic and cultural context of the nations being aided. Many well-intentioned efforts have been misdirected. At best, such efforts do no harm beyond wasting resources; at times, such efforts have left many in the recipient nations in worse conditions (Devarajan, Dollar, & Holmgren, 2000). In the area of health care, a basic problem is assuming that the residents of developing nations need advanced medical services. Unfortunately, developing nations cannot support advanced technologies, and the foreign aid available can provide this type of care to only a small minority of the population.

Bringing foreign nationals to European or North American medical schools for training means that such people will become accustomed to technology-intensive medical practices. Partially as a result of this training, some physicians have returned to their home countries to practice in city hospitals serving only the rich; in recent years 90% of the medical resources of Nigeria have been spent in cities where only a minority of the population lives (Manoff, 1985). This happens while many rural children go without immunizations that are inexpensive. Some have said that much foreign aid has actually harmed the recipient nations (Easterly, 2007).

A favorable outcome occurred when planners examined treatment for severe diarrhea, which can lead to the death of children in poor areas if the fluids and nutrients lost are not replaced. Such fluids can be replaced using intravenous feeding; however, the cost would be staggering for an undeveloped nation. Premixed packs of salt, sugars, and other ingredients were developed to make it easier to prepare a drink for affected children. Widespread introduction of such premixed packages, although costing only a fraction of intravenous methods, would have been unfortunate because the packs were still too expensive to supply everyone in need and their use would imply that the people could not take care of themselves. Manoff (1985) reported that planners calculated that 750 million packs would have been needed in 1990. A better approach was found: locally available ingredients were used in recipes that could be taught to residents. Health care programs that provide medically appropriate skills are far more effective than ones that meet the needs of a few but foster feelings of helplessness among many others.

Failing to Relate Need to Implementation Plans

Sometimes a program is prepared on the basis of a reasonable understanding of unmet needs, but the implementation makes it difficult for people to participate. Food stamps are designed to assist poor families to purchase food at reduced costs; it is a compassionate response to nutrition needs. Unfortunately, through local indifference and target group ignorance only 20% of those eligible for food stamps were using them in the early years of the program. An advertising campaign in New Mexico tripled the number of eligible people using food stamps in six months (Manoff, 1985). In 1997 only half of eligible households participated in the food stamp program (Castner & Cody, 1999). Some health promotion programs fail to consider all aspects of effective marketing (Winett, 1995).

Failing to Deal with Ignorance of Need

If a group in need does not recognize the need, a program must include an educational effort (Cagle & Banks, 1986; Conner, Jacobi, Altman, & Aslanian, 1985). People screening for and treating high blood pressure deal daily with patients who do not recognize a need for care. When the condition first develops, elevated blood pressure seldom produces noticeable symptoms; in addition, medication can have undesirable side effects (Testa, Anderson, Nackley, & Hollenberg, 1993). Because patients expect to feel better, not worse, after taking medications, many reject the treatment. A rejected treatment means program failure, even though the medication is effective in reducing the likelihood of strokes and other cardiovascular diseases.

Some people believe that when it comes to good health practices and the availability of educational opportunities, society fulfills its responsibility by simply providing the facts. They argue that publicly funded service providers are under no obligation to persuade anyone to take advantage of the programs. By contrast, social marketing techniques (Weinreich, 1999) are used to foster the adoption of more adaptive health behaviors, wise use of government services, and the maintenance of safer environments. Social marketing is needed not only to provide basic education, but to strengthen motivation. Such efforts are needed to counter the extensive advertising that encourages the use of products that can harm health, such as very sweet and high-fat foods, cigarettes, and alcohol. The city of Cleveland launched a social marketing effort beginning in 2008 to encourage children to value education and to counter "gangsta rap posters" and other negative influences (Perkins, 2008).

Using Needs Assessments in Program Planning

Once the level of need has been assessed for a particular population, then program planning may begin. Planners seek to develop a service or intervention to help the population achieve or approach a satisfactory state. The probability of developing a successful program increases markedly when the important stakeholder groups are involved in the planning and cooperate in selecting both the services to be offered and the mechanisms for delivering them. Stakeholders can be involved through formal representation on committees or in focus groups responding to developing plans.

After outcome goals have been specified, the next step is to consider the intermediate goals that need to be achieved on the way to achieving outcome goals. Intermediate goals could include, for example,

accomplishments, behaviors, attitudes, or levels of knowledge. In order to specify intermediate goals that would indicate how much the target population is being helped, planners must describe the program theory. Unfortunately, as mentioned in Chapter 2, planners sometimes develop programs without asking what must happen to the program participants before they can reach the outcomes hoped for; in other words, an impact model ought to be developed as the program is planned.

After developing an impact model, the planning group can specify what actions the agency must take to implement the program. The resources that need to be considered include the knowledge and skill of the staff, the physical resources, financial resources, advertising, and outreach efforts. Since planners cannot operate with blank-check assumptions, the projected resources required are compared with the financial support that would be available. If the resources are not sufficient, the planners have several options. One temptation is to leave the goals intact while reducing the intensity of the program and hope for the best. If the planning had been sound and if the conceptual basis of the plans is valid, such a decision is a prescription for failure (Rivlin, 1990). To fail in the complex task of helping people change is not a disgrace, but to continue to plan a program that planners suspect cannot work well because it is too weak is a waste of valuable resources and fosters cynicism about social programs. More useful alternatives include reducing the number of people to be served, narrowing the focus of the service, or even changing the plans to meet a different need, one that requires fewer resources.

Although this presentation seems straightforward and linear, the planning committee will frequently check back to verify that they have been faithful to the mission of the organization that will sponsor the program. In fact, while developing intermediate goals, planners might discover that some of the outcome goals were unrealistic. What does not change is the need to work in a generally backward direction during the planning process, beginning with what is to be accomplished and ending with a plan to achieve those goals (see Egan, 1988a,b). If there are no clear goals, rational planning cannot occur. Surprisingly often, people in agencies and organizations sense a need, but do not measure it carefully or even explore it in a qualitative manner (Mathison, 1994). Sometimes they then watch helplessly as the program fails to have an observable impact on the population in need. Altschuld and Witkin (2000) discuss how evaluators can help in translating a needs assessment into a program, but the actual planning is not a primary responsibility of evaluators.

CASE STUDY 6

Improving Systems to Address Needs

Tony L. Sheppard, PsyD, CGP, FAGPA was involved in the follow-up phase of a program evaluation project while on his predoctoral internship with a large community mental health agency in the Midwest in 2001–2002, building on the initial phase started by the previous predoctoral internship class. The community mental health center, recognized for its substance abuse programs as well as for those treating chronic mental illness, wanted to better understand how their services addressed the comorbidities among these presenting problems. As a part of the needs assessment phase of the study, it was discovered that the diagnosis of psychiatric conditions in their addiction treatment programs and the diagnosis of substance abuse disorders in their psychiatric programs were both well below the state and national averages. By interviewing staff, reviewing medical record data, and considering agency forms and paperwork, it appeared that the staff in the psychiatric programs were not screening for substance abuse adequately and that the staff in the substance abuse programs were not screening adequately for psychiatric conditions. Further, the paperwork used in the respective programs did not facilitate such screening.

The initial evaluation recommended cross training. Staff from the substance abuse programs would train those in the psychiatric programs in screening for and diagnosing addictions. Likewise, the staff in the psychiatric programs would train those in the addiction programs in screening for and diagnosing psychiatric conditions. Further, staff from both programs would collaborate on ensuring that paperwork facilitated this process. In some cases, this was as simple as including substance use screening questions and psychiatric symptom checklists on intake forms.

Meeting Needs

Although many needs assessments are conducted before a program exists, often to determine whether or not to start the program, this case illustrates the value of needs assessment in an established program. As noted, the agency intended to serve a substantial number of people with comorbid substance abuse and other chronic mental illness, knowing that many of those with both problems seek services for one or the other, but not both. Despite this good intention, various factors interfered with their goals, especially clearly identifying the clients with comorbidities. This case illustrates that the various elements of program evaluation are much more integrated than separate, as understanding the actual condition of the clients who sought their services involved also understanding the actual practices of the agency (implementation) in detail.

Implementation

This case highlights an important point for evaluators—challenges within programs and the corresponding opportunities for improvement do not always or even generally result from avoidance or resistance in implementation. The agency and the staff specifically wanted to identify clients who would benefit from the coordinated treatment of both substance abuse and other psychiatric disorders. The fact that substantially fewer clients than expected were recognized as having comorbidities was identified by careful examination of multiple sources of information and the discrepancy between agency diagnoses and external norms. Although not explicitly stated in the evaluation, it is very likely that a more engaging style of evaluation made it possible to discover the problem, whereas evaluators with a style that staff found punitive or threatening might well never have learned the essential details.

Stakeholders

Clearly the clients and staff were important stakeholders. The clients of this community mental health center were often from underprivileged populations. Their awareness of comorbidity and therefore, their ability to advocate for themselves were generally lower than in more privileged groups. The staff and that agency as a whole are also important stakeholders. This agency is consistently recognized as a regional leader in innovative and effective treatment. It is vital to the reputation of the organization to maintain the highest standards. Further, the agency prides itself on providing its staff the tools necessary to perform their jobs proficiently.

Side Effects

Although not explicitly noted as a side effect, the factors that prevented initial determination of comorbidity were clearly a negative unintended consequence of the way the program was originally conducted. No positive side effects were noted.

Improvement Focus

As implied in the Implementation section, the genuine interest of the evaluators to help the program not simply improve but to meet their desired goals was an obvious component. Not to overemphasize the point, but evaluators who assume good intentions on the part of program stakeholders generally develop greater trust, better cooperation, and more useful findings. Further, good evaluators also know that having their own good intentions is most helpfully matched by communicating those good intentions consistently to stakeholders.

Outcomes

Following the cross-training and paperwork revision processes, agency records were reviewed again and staff were interviewed regarding their perceived proficiency in assessing for the given symptoms. Not only did staff report improved proficiency in the areas addressed, the data suggested that comorbidities were more in line with the state and national numbers. This program evaluation project was written up and shared agency-wide as an example of quality improvement. This was used to build morale in the agency, but also to encourage the rural clinics in the organization to embrace these findings and, likewise, adopt the new forms and procedures.

Nuances/Mechanisms

As noted previously, this project was shared agency-wide as an example of quality improvement. This particular agency prides itself on this type of dedication to its employees and clients. This process was shared with the rural clinics that often have the same clinicians providing both services. The paperwork revisions were shared with the rural clinics enabling them to better screen for comorbid conditions.

Summary

Involvement in this project inspired Dr. Sheppard to remain involved in ongoing program monitoring and improvement. It demonstrates a commitment to continual monitoring of one's processes to ensure that the best possible services are offered. It also emphasizes the importance of accurate diagnosis in the provision of services in a large health system.

Source: T. L. Sheppard (personal communication, January 16, 2018).

EVALUATOR PROFILE 6

Tony L. Sheppard, PsyD, Local Clinical Scientist Evaluator

Dr. Sheppard came into the field of clinical psychology in good company—like many others, he understood the work "affects people's lives," and he wanted to be a part of that kind of valuable endeavor. More specifically, however, he strongly connected with the philosophy of his doctoral program to train "Local Clinical Scientists" (Stricker & Trierweiler, 1995) who reflect systematically on their practice, connecting their experiences to the research in the field, and learn to do even better with the next case. He recalls that when he took the program evaluation class with Dr. Andy Meyer, he said, "I want program evaluation to be a part of my career," and he chose his internship site (see Case Study 6) because of its focus on program evaluation. He currently owns his own private practice that emphasizes his passion for high quality group therapy, especially for children and adolescents, in which he is "constantly engaging" in some level of program

evaluation—ranging from some elements of it to much more formal processes.

One clear example of his work to blend science and clinical practice in the Local Clinical Scientist model is finding the appropriate level of rigor to address challenges. With his reputation regarding group therapy and with multiple groups at his practice available, his doctoral clinical psychology practicum students commonly field phone calls from parents who want to know if their child can join one of the groups. Tony described the dilemma of how best to respond to these calls. On the one hand, helping clients find the right group is vital, and he is committed to doing the best he can, within reason, to find good matches. But, on the other hand, making good matches require knowing details about a client's case, and the most rigorous way to determine those elements would be with a formal assessment, which is time-consuming and expensive. While he values the precise insights that can be gained, he worries that too many potential clients would take the time and pay the high price only to be told, "None of our groups would be the best fit for you, try these other places."

His solution has been to develop a script for his students to follow during these phone call inquiries, assessing sufficient details that he has confidence in the plan. Although many clinicians might describe a similar set of instructions for office staff, Dr. Sheppard developed his script using program evaluation principles and practices to assure that the results were good enough for his high standards. He did not rely just on his clinical training; he also incorporated systematic evaluation to achieve a better process. Or, as he puts it, he works to be the Local Clinical Scientist version of a program evaluator.

Sources: T. L. Sheppard (personal communication, February 2, 2018); Stricker & Trierweiler, 1995.

SUMMARY AND PREVIEW

The importance of assessing needs before beginning to plan programs cannot be overstated. The definition of need emphasized that evaluators try to learn what people need in order to be in a satisfactory state, in their social context. Although we tend to know a lot about our own needs, we often do not fully understand what it takes to satisfy those needs, as illustrated by marginal understanding of objective nutritional needs. Since educational, psychological, and social needs are very complex, it would not be surprising to learn that these needs are not fully understood either. Consequently, it is wise to seek information on needs from a variety of sources. Community surveys, agencies currently serving people, social indicators, focus groups, and expert informants are sources of information on needs. Once needs are clear, planning may begin. Planning has the best chance of success if the conceptual basis of the program is developed and used to set the outcome and intermediate goals.

After programs are in place, evaluators continue to contribute since programs must be monitored to guarantee their integrity. Furthermore, information gathered by monitoring serves to show what aspects of a program should be adjusted. Chapter 7 includes an introduction to management information systems and their use in program evaluation.

STUDY QUESTIONS

1. Examine the daily newspapers for articles on social services or consumer products that show how a human service or a new product was affected by an analysis of need. Note that if the example concerns a consumer product, the definition of unmet need in this chapter may not be appropriate; in such cases, manufacturers are usually dealing with meeting wishes or preferences, not serving unmet needs.
2. Look for examples of agencies or producers who failed because they did not conduct an appropriate needs assessment before offering a service or bringing a product to the marketplace. What kind of information might the organization have sought to avoid the error?
3. Many colleges sponsor a learning assistance program for students who do not perform as well as they wish in their classes. Sketch a plan to do an assessment of needs for such a program. Assume that there is a program already, but the college dean wants to know how well the services match the needs and whether services should be expanded. Consider the information you want and the sources as well as the methodology to use in gathering the information.
4. Under what conditions would an assessment of needs be threatening to an organization?
5. Thinking of unmet needs sometimes clarifies planning. Sometimes we hear that a community needs something, a recreation center, for example. What unmet needs might underlie the assertion that a town needs a recreation center? What would you measure to test that idea? What are the alternative ways of meeting the unmet need you thought of?

ADDITIONAL RESOURCE

Altschuld, J. W. (Ed.) (2010). *The needs assessment kit*. Thousand Oaks, CA: Sage.

This is a set of five books addressing needs assessment. The five are: Book 1, *Needs Assessment: An Overview*; Book 2, *Needs Assessment Phase I: Getting Started*; Book 3, *Needs Assessment Phase II: Collecting Data*; Book 4, *Needs Assessment: Analysis and Prioritization*; and Book 5, *Needs Assessment Phase III: Taking Action for Change*. It is a comprehensive guide that provides more information than many evaluations will need, but most evaluators will eventually run into situations or questions that are addressed well in this series.

7 Monitoring the Implementation and the Operation of Programs

THE STRUCTURE OF THIS CHAPTER

As you have already seen, this edition uses the many advantages of online materials as an important component of the textbook. You have occasionally been directed to the eResources for a few elements that, because of size or other characteristics, are better online than in print. In the case of one of the central elements of this chapter—information systems—the usual pattern of having online materials supplement print is reversed. Learning about ways to enter information into any current system, to manage that information, and then to use that information for any of a number of elements like monitoring the program, reporting on outcomes, and improving the program's efficiency virtually requires a dynamic interaction with computer systems rather than just reading text or viewing small, static images.

So, this chapter begins in the usual way—describing the element of program evaluations known as monitoring the implementation, providing a good overview of the area and a general introduction to the elements of it. And the chapter concludes in the usual way, describing potential problems to avoid, before providing a relevant case study that especially illustrates the theme of this chapter, followed by a profile, summary, and questions. But for a substantial portion of the content of this chapter, you are directed to the eResources to learn about information systems with materials that are a better match to what is involved in using those systems than can be provided in a single chapter of a printed textbook. Among other things, a common request to program evaluators currently is for advice regarding what systems an agency might purchase to manage their data. Although the eResources module will not make you an information systems expert, you will gain enough exposure to the relevant points to be able to guide clients in a responsible way. Further, technology is changing so quickly that substantial developments are likely in such a short time that printed information will quickly become more and more dated, whereas the eResources will be updated on at least a yearly basis.

INTRODUCTION TO MONITORING

A classic statement that program monitoring "is the least acknowledged but probably most practiced category of evaluation" (ERS Standards Committee, 1982, p. 10) is still relevant today. The most fundamental form of program evaluation is an examination of the program itself—its activities, the population it serves, how it functions, and the condition of its participants. Program monitoring includes an assessment of the resources devoted to the program and whether this effort is implemented as planned. It is a serious mistake to plan an evaluation that focuses only on the long-term results, ignoring its activities. Careful program monitoring yields "impressive results" (Lipsey, Crosse, Dunkle, Pollard, & Stobart, 1985).

Monitoring is a basic activity of any organization. Businesses monitor their performance by counting receipts at the end of every day and by performing inventories of stock. Human service managers monitor their activities in some ways similar to those used by for-profit businesses. However, education, health care, criminal justice, and welfare agencies are designed to change the people served in some way, not simply to please them, a critical difference (Weiss, 2002). Consequently, the end products of human services—well-functioning people—cannot be measured as reliably, validly, or easily as store inventories or account balances. The place to begin is to ask whether a program has been implemented as planned. If messages showing why using cell phones while driving distracts drivers are not published or broadcast, there is no reason to evaluate

the impact of the material. If drug abusers do not attend a drug rehabilitation program, then there is no need to ask whether the program reduced drug abuse.

Managers need basic information to facilitate their planning and decision making, to anticipate problems, to verify that program plans have been carried out, and to justify continued support for the program. If it is discovered that the people receiving the services of a program have unmet needs that are quite different from the needs a program was planned to meet, the program must be altered. An increase in the participant dropout rate would suggest problems. Without a program-monitoring system, managers might be slow in noticing the problem until it is severe and difficult to correct. The techniques described in this chapter apply especially to programs expected to remain in place. Many programs such as college counseling centers, community mental health centers, and organizations serving the needs of homeless people are expected to continue to offer services for years; these organizations need convenient access to data describing the service.

Evaluators have shown the value of monitoring the process of providing a product or service, not just evaluating the final outcome (Mowbray, Bybee, Holter, & Lewandowski, 2006). A management information system provides information showing that the program is operating as planned. Keeping track of operations is similar to following progress during an automobile trip as we watch for route markers and familiar landmarks. While planning a trip, most drivers take note of intermediate locations in order to monitor their progress and to be sure they stay on the correct route. We don't simply plan a trip, aim the car, drive for four hours, and ask, "Are we there?" The last point is so obvious that it is hard to understand why so many service providers, program planners, and evaluators fail to apply this principle to program evaluation. All too often an evaluation plan includes only an assessment of success in achieving the ultimate outcome, not the services actually provided or the intermediate steps needed to reach the desired outcome (Gehrs et al., 2004; Lipsey, Petrie, Weisburd, & Gottfredson, 2006). Indeed, it remains true that some program sponsors do not even think about an evaluation until a program is well underway.

MONITORING AS A MEANS OF EVALUATING PROGRAMS

Some of the information that evaluators summarize for program staff are fairly straightforward descriptions of the clients served and the services provided. At times, managers and staff seem to believe that such information is unnecessary because they think that they have an accurate and complete understanding of the service they provide. However, it is common knowledge among evaluators that managers and staff often hold opinions that are incorrect. Agency counselors overestimated the median number of treatment sessions clients came to (Posavac & Hartung, 1977; Pulford, Adams, & Sheridan, 2008). Solberg et al. (1997) described the need for better information systems in medical care, commenting that "… most of the defects in the current process of care for patients with diabetes or other chronic diseases are the result of an inconsistent, disorganized approach *that is the natural result* of relying almost entirely on the memories of individual clinicians" (1997, p. 586, italics in original). When information systems have been implemented, improvements in care have resulted (Ricciardi & Tabb, 2006; Williams, Schmaltz, Morton, Koss, & Loeb, 2005). Because it is so easy to be misled by our impressions, a summary of the client population and the services provided would be an appropriate first step to demonstrate effective program implementation and to facilitate effective management.

A description of clients and services can be done once as a snapshot of the agency or as a regular activity in order to track the agency over time. Managers of smaller and newer programs may benefit from summaries of information in the records; this process is time-consuming and sensitive because files contain far more material than would be needed to describe the agency's clients and the services provided. Larger programs usually need periodic reports for accrediting and regulatory agencies, each with its unique information requirements. A procedure to gather information regularly is called a management information system. The system can be used at any time by staff. Evaluators can periodically summarize patterns and trends to track the activity of the agency. Further, many agencies have already instituted electronic medical records systems.

WHAT TO SUMMARIZE WITH INFORMATION SYSTEMS

Relevant Information

The information gathered must be central to the purpose and mission of the program. The appropriate information can help set staff levels, satisfy accreditation standards, plan for space needs, and monitor quality. Effective evaluators gather only information that is important, not information that might be simply interesting. Because every piece of information recorded increases the costs of data collection and analysis, evaluators guard against including information that is not of use to the agency.

Actual State of Program

An information system must describe the actual state of the current program. Evaluators are careful to distinguish between the program as described and the program as administered. Many programs that look very good on paper fail because staff did not or could not follow the design or procedure as planned. Reporters visited Chicago city parks during periods when activities were officially scheduled, only to find field houses locked or abandoned (Stein & Recktenwald, 1990). If the reporters had limited their study to the official activity schedules, they would have assumed that the programs were being offered. There is no excuse for an evaluator to permit written plans, no matter how official, to substitute for on-site observations.

The information necessary to describe the services rendered includes (1) the type of services offered by the program (such as job training, group therapy, immunizations), (2) the extensiveness of the program (such as hours of training or therapy, skills to be learned, facilities), (3) the number of people participating in the service, especially the proportion of people completing education or therapeutic programs, and (4) indicators of quality. To be useful the system must permit convenient retrieval and summarization of information. Expected numbers of program participants are usually included in plans; such numbers could be used as benchmarks, providing a clear indication whether implementation faithfully reflected the plans.

Program Participants

If a program is well planned, the population to be served would have been specified carefully and the program tailored to meet the unmet needs. An information system should document the identity and the needs of those using the service and compare these findings to the program mission. In other words, the fit between the program and the needs of those using it should be examined. If the fit is not close, changes in the program can be considered. If the agency had not defined the group to be served, a description of the people using the agency may be especially enlightening.

Program participants are routinely described by gender and age. The choice of additional information depends on the specific nature of the program. Such information may include where the people live, the major problems for which they are seeking assistance, the source of referral, and ethnic background. In some settings the barriers to obtaining services might be important to know. Transportation difficulties, childcare needs, and length of time spent on the waiting list might be useful in describing a program and suggesting improvements.

Providers of Services

The information system should include a description of who gives the services. In some settings, less-well-trained people work under someone else; a casual examination of the program might imply that the supervisor was the provider. For example, a licensed psychiatrist may sign insurance forms, but employ social workers to provide therapy. To understand a program knowing who actually works with the participants would be essential. Staff may be underqualified or overqualified.

PROGRAM RECORDS AND INFORMATION SYSTEMS

When appropriate information is gathered consistently, the types of reports illustrated in this chapter can be prepared without great cost.

Problems with Agency Records

It is impossible to monitor a program that does not keep adequate records (Vroom, Colombo, & Nahan, 1994). However, it is no secret that the records of many human services are often abysmally incomplete. Lurigio and Swartz (1994) described the difficulties they had as they tried to work with the inadequate records of jail-based drug treatment programs. Records are often in poor condition because recordkeeping is dull and does take time. Individuals providing human services often could provide more services than they have time to do and, thus, view keeping records as time taken from clients, students, or patients. Records were less important when impressionistic evaluations of success and effort were sufficient. However, impressionistic evaluations no longer satisfy the requirements of regulatory and accrediting agencies. It is critical to maintain and summarize agency records.

Increasing the Usefulness of Records

Methods of summarizing information must be developed to increase the usefulness of files because information that cannot be summarized is not helpful to evaluators or program stakeholders. Information systems permit the regular summary of participant descriptors, services provided, and quality indicators without manually searching through files.

When planning an information system, the program director, staff, and other information users develop a list of essential variables useful in managing and evaluating the program. A point repeatedly stressed in this book is that information is gathered only when it is useful. If there are legal requirements to obtain certain information, or if decisions are to be made on the basis of information, then it should be gathered. If material could reveal suspected negative side effects, it should be included as well. The challenge to information system developers is to summarize such information in helpful ways. Managers and staff need relevant and useful information, not just lots of details (Ackoff, 1986).

At this point, proceed with the "Information Systems" module on eResources, and then return to this point.

AVOIDING COMMON PROBLEMS IN IMPLEMENTING AN INFORMATION SYSTEM

Guiding staff on their use of an information system is a major effort; there are a number of possible pitfalls in the process.

Guard against the Misuse of the Information

As with all information, a report comparing individuals with each other can be misused. Critics fear that providing summaries of information usually hidden in files allows such data to be used in a vindictive manner; however, most evaluators believe that providing more information reduces the opportunities for directors to behave autocratically. Most unfair managers restrict information, just as despotic rulers control the media in the nations they control. Nevertheless, it is crucial to remove fear from an evaluation system. When people fear that information will be used against them, they are motivated to hide their errors (Campbell, 1987; Deming, 1986; Edmondson, 1996, 2004; Hilkevitch, 2000) and to avoid asking for assistance.

Managers must also guard against the possibility that acceptable variations in style will appear as deficiencies. Some therapists may encourage clients to leave therapy sooner than others do. They may do this because of theoretical convictions, not because of inadequate therapeutic skills. In the table on page 16 of the eResources module for this chapter, this confusion was avoided by reporting separately the number of clients quitting after less than three counseling sessions. Even therapists who plan on short-term therapy expect to serve clients beyond two sessions. A related danger is the temptation to use the system in a misguided attempt to monitor staff work at such a detailed level that staff members simply report fictitious information.

Avoid Setting Arbitrary Standards

It is critical to avoid setting arbitrary standards for the staff to meet. The indexes developed using information systems can indeed reflect quality services, but they reflect only some aspects of the service. If pay and

promotions of staff are based on these partial views of quality, staff will find ways to meet the standards at the expense of other aspects of their work. Primary and secondary school teachers have learned how to structure lessons to raise standardized test scores that are thought to represent the outcomes of effective schools, and college teachers know that assigning higher grades increases the teaching ratings made by students (Gillmore & Greenwald, 1999). Campbell warned that focusing on a particular variable would increase the probability that it would become invalid, indeed corrupted (see Shadish, Cook, & Leviton, 1991, p. 141). Management information is used most effectively when combined with additional measures of quality.

Avoid Serving the Needs of Only One Group

People filling different roles in an organization have different informational needs because their responsibilities differ. The needs of the accounting department differ from those of the billing department, which in turn contrast with those of the service delivery staff. By working closely with only one or two such groups, evaluators run the risk of providing information useful only to some groups. If that happens, the information system would be seen as irrelevant by those whose needs remain unmet by the system; such people will resist providing information for the system, thus reducing its usefulness.

Avoid Duplicating Records

Human service staff will not cooperate happily with an information system that duplicates information already reported and stored elsewhere. It is undesirable to simply impose an information system as an addition to existing recording procedures; integrating the information required by the old system with the automation of the new system is preferable. Staff members may question the accuracy of the new system. Because staff can seldom describe the population they serve in quantitative terms, they will sometimes be surprised at the summaries from a management information system. As the validity of the system is demonstrated, staff may come to appreciate the added knowledge provided by an information system.

Avoid Adding to the Work of the Staff

An integrated information system was developed in an agency providing family and marriage counseling (Savaya, 1998). Although the staff members were consulted in development of the system and were given training in its use, they were expected to complete assessment forms on their own time. Furthermore, the counseling staff was expected to change the way they assessed the success of their clients. The limited cooperation of the staff made it impossible to use the information system effectively. In contrast, to develop the most useful system, the users must be involved in its design. The information needed, the manner of summarizing it, the frequency of reports, and other issues should be determined in cooperation with those who are expected to use the information. The evaluator may be enthusiastic about getting the system operating; however, learning what is useful requires patient work with the staff and administrators.

Avoid Obsession with Technology

Although almost no program can be run successfully in these times without electronic records at some level, it is important to find the balance between using technology as a very helpful set of tools and also maintaining the human connection that people desire. When the use of technology has a clear benefit of savings of time or money, it may be obvious to clients that they share in those benefits. For example, if a survey to document the level of symptoms before treatment is e-mailed to clients to save the time of completing a lengthy survey in the office, many clients will appreciate the convenience of being able to answer somewhat sensitive questions on their own schedule and in the privacy of their home. Further, if a comment indicates that those clients who want to complete a hard copy survey at the agency may do so, any who are not comfortable with the online version will not feel forced to complete it. But when the use of technology functions only to keep clients at a distance, any benefits are overshadowed by the obvious problems with such an approach. Although the line between these two areas certainly changes over time, the main point is that paying attention to the message conveyed in the use of technology—whether it helps the connection between clients and the agency or serves

only to widen the gap between them—is an important element to consider.

Case Study 7 reports on the program evaluation of a federally funded program that focused on the health of mothers and infants. The program had been going on for several years and was in the second year of the grant cycle when Data Driven Solutions, LLC (DDS) was retained by a local governmental agency for the evaluation. As you will see, although the evaluators initially took the approach of a general program evaluation, issues of implementation quickly became central, and the section on implementation takes center stage in the case study. Some evaluations begin with a focus on implementation, whereas others develop it in response to the specific program.

CASE STUDY 7

Adapting to the Needs of the Program

The program addressed the health of mothers and infants through education on issues like healthy behaviors, breastfeeding, and safe sleep practices for infants; referrals; and support for prenatal and postpartum women.

Meeting Needs

Although everyone has a general need for good health, expectant and new mothers as well as infants have particular vulnerabilities to threats to their health, and these are substantially greater in vulnerable populations such as those with low income. In this particular case, the granter did not require extensive documentation of the details of these needs, so general recognition of these needs was sufficient.

Implementation

As part of the program, a large amount of data was collected on each of the participants of the program, including demographic data, background data, and data to address specific benchmarks of interest to the federal government, such as percentage of fatherhood involvement and percentage of participants with health insurance. When the program originated, a database was developed and data were entered from paper records of visits and meetings with the participants. The structure of the database supported the specific reports required by the federal government and those reports were easily generated.

The evaluators quickly discovered, however, that a series of changes, many of which were relatively minor and could be addressed individually, resulted in a cumulative effect that completely transformed the report process. First, additional fields had been added at various times to the database. In itself, this change was neither unexpected nor a problem. But next, some of the forms used to collect the data changed and the process of data entry also changed. Again, each individual change was not a major problem, but the series of shifts meant that the likelihood of errors and other problems began to increase more quickly. As readers might imagine, when paper forms and computer data fields are arranged in a similar layout, it is fairly easy to connect the right written information with the right computer entry. When new fields are added and then the forms are changed so both the words and the arrangement are different, it can be much easier to enter a number in the wrong place.

In many ways the "final straw" came when new benchmarks were established and yet another new set of forms were required. At this point, an entirely new database was needed to match the other elements in a way that supported efficient and accurate recordkeeping, rather than trying to add more layers of fixes to previous layers. Although one aspect of implementation, the nature of the actual services, had not been directly affected in a major way by these many changes, documentation of the services and the way that relevant data regarding those served were collected, a second aspect of implementation, was completely transformed.

To complicate the evaluation even more, however, a final report using data in the old database system was still required. At this point, the results provided by the outdated automated formulas in the old system seemed blatantly inaccurate to program staff, and they asked the evaluators to examine these findings. A detailed inspection of the old database found problems of missing data, obvious errors such as impossible values for given measures, and related problems. Some of the errors could be corrected by reviewing other records or tracing information entry backwards, whereas in other cases, it was clear that data had been lost, and those missing cases were ignored. The results calculated after this lengthy process appeared to both staff and funders as sufficiently accurate under the circumstances, which was a very successful outcome.

As noted in the text, one form of program evaluation is an "implementation evaluation" that is structured primarily to assess how a program has been initiated, to document what has actually been implemented, and to set the stage for a later accounting of outcomes. In this case, the original intent of a general evaluation was modified due to the overwhelming need to address the recordkeeping aspect of implementation, although some elements of a broader evaluation remained.

Stakeholders

Main stakeholders are noted as the mothers, the infants, the healthcare community, and the federal program that sponsored the services. In this case, adjusting the reporting system to the needs of the program while preventing the adjustments from interfering with continuing service to the clients was a substantial challenge.

Side Effects

Especially with the focus on the data system, no side effects were noted.

Improvement Focus

On a general scale, the focus on improvement should be fairly obvious—adapting details of the evaluation to fit the needs in an appropriate way and finding a process to facilitate changes in the organization to meet the needs of the funder. Hopefully it is clear that adjustments to an evaluation plan can be an excellent way to address broader evaluation needs by helping the program improve for the future. It is also true that some adjustments are inappropriate, such as changing the criteria for success to make poor results look better for staff. Understanding the difference between the two types of adjustments requires good judgment and integrity on the part of evaluators.

Outcomes

In this case, because the funding agency was most interested in reports about meeting benchmarks, the outcomes could be addressed appropriately even with data that were fewer and less complete than ideal. As with the above point about using good judgment to make decisions about adjusting an evaluation plan, knowing when one report in a series of reports can appropriately address a transition involves using good judgment.

Nuances

Nuances were not explicitly addressed.

Summary

As this case illustrates, the job of an evaluator may involve much more than determining whether a program is meeting its expected outputs and outcomes. In this situation, the evaluators worked on insuring all data points were included in the final results by developing an in-depth understanding of the database that contained all the raw data.

Source: M. Dever (personal communication, January 25, 2018).

EVALUATOR PROFILE 7

Marie Dever, JD, MPH

Marie earned a Master of Public Health from the University of Louisville School of Public Health and Information Sciences in May 2015. She was inducted into the Delta Omega Honorary Society of Public Health, Beta Pi Chapter. While earning her degree, Marie concentrated in biostatistics, taking classes in biostatistics, regression, and survey methods. She has utilized the data analysis programs SPSS, SAS, and R. In addition, she has taken classes in GIS (geographic information systems) and program evaluation. She completed a practicum at the University of Louisville Division of Infectious Disease working with refugee health data. Finally, Marie became certified in public health in May 2015.

Since 1994, Marie has been a licensed attorney in the Commonwealth of Kentucky and until June 2015, she was a practicing attorney. She started her career at the Jefferson District Public Defenders District Office. After working there for five years, she spent the rest of her legal career specializing in litigation, insurance defense, medical malpractice defense, and long-term care facility defense.

Marie brings a unique perspective to the field of data gathering, management, and analysis as a result of her education, training, and experience. Due to her legal experience, she has excellent writing, public speaking, problem solving, and analytical skills. Her education and experience in the public health field has provided her with an understanding of the challenges and needs of public health agencies and organizations. Marie's singular background means that she can offer invaluable assistance to any agencies or organizations needing assistance with data gathering, management, and analysis.

Source: M. Dever (personal communication, January 22, 2018).

SUMMARY AND PREVIEW

Human service agencies can begin an evaluation with a systematic description of their programs, the amount and type of services provided, and the identity of those receiving the services. Such monitoring permits an agency to account for the funds expended and makes it easier to compare a program as implemented with the original plans. These sorts of comparisons can be invaluable in showing what aspect of a program needs additional effort or why a program never had the impact envisioned. Also, the need to complete accrediting agency and government surveys can be met more easily if monitoring for quality assurance is carried out regularly rather than only after requests for information are made.

Monitoring is a crucial aspect of program evaluation; it is essential to keep track of activities. However, stakeholders want to know how successful programs are in helping participants, students, or patients. The

next chapters focus on methods that evaluators and stakeholders use to examine the degree to which outcomes are achieved. Such evaluations would be used with reports from the information system for a more complete understanding of the agency.

STUDY QUESTIONS

1 How does program monitoring fit into the schematic diagram of the function of program evaluation given in Figure 1.1?
2 Explain how monitoring relates to formative and summative evaluations.
3 People sometimes object to an evaluation of effort on the grounds that monitoring activities do not focus directly on the quality of the services provided. Why does this criticism not negate the reasons for monitoring the activities making up the program?
4 Suppose that you managed a telephone sales office. Twenty people make telephone calls to teachers who use educational DVDs. If these potential buyers are interested in previewing a DVD, they are given a code to permit them to preview it online for seven days to decide whether their school would buy it. How would you develop a monitoring system to keep track of what your employees are doing? How would you measure their success? What type of indexes of success would be important?
5 What types of program evaluation questions can be answered using an information system? What are some questions that an information system cannot answer?

ADDITIONAL RESOURCE

Along the lines that information systems themselves are changing so rapidly that the eResources support timely updates, the best resources to supplement this chapter are also changing rapidly. See the recommendations in the eResources for print and other forms of materials for further reading and experiences.

8 Qualitative Evaluation Methods

All approaches to program evaluation noted in Chapter 2 and in the eResources have advantages and disadvantages. One of the major advantages of qualitative methods is its flexibility and appropriate use in innovative and novel settings. When stakeholders want information quickly or want to understand a complex program with ill-defined (even conflicting) objectives, highly structured approaches to program evaluation may not be appropriate. Imagine that a new president of a private college has become concerned about the proportion of first-year students who, although in good academic standing, do not register for the sophomore year. The president responds to this problem by proposing the development of curriculum changes, enhancements to available student activities, dormitory reorganizations, increased access to computer terminals, and other steps to improve the experiences of first-year students. It is hoped that such measures would lead more first-year students to identify with the college, continue their education there rather than transferring, and, in time, become alumni who participate in fundraising campaigns.

After a period of planning and negotiation among the president's office, student life administrators, a skeptical faculty, admissions, and student organizations, plans are made and the program is implemented. The stakeholders would want timely information to help them adjust to these innovations. Some aspects probably need to be strengthened and others perhaps should be dropped. Since a lot of effort and funds are devoted to the program, it should be evaluated. But how?

The program needs to be evaluated, because without an informed evaluation, cynical stakeholders will claim evidence for failure, although others may see evidence for success. The question is how to evaluate it fairly and in a way that provides useful information for the stakeholders. The president's office should have a central administrator keeping track of the activities sponsored by the innovation (the outputs), but more is needed because stakeholders want to know more than an information system can provide. Qualitative evaluators would supplement the numerical summaries from a relational database with direct observations of activities associated with the program; discussions with participants, with those who chose not to participate, and with staff; and examinations of program documents, other materials, and artifacts. Often this approach is called *qualitative evaluation* to distinguish it from quantitative measurement.

The term *qualitative data* is used in several ways. Some people think of questionnaire items referring to ratings of impressions or subjective reactions as qualitative data; this is not how the term is used here. Such ratings are numerical and are treated as quantitative data. Others think of the characteristics of people (e.g., college major, on-campus vs. off-campus residence, religious affiliation, etc.) as qualitative variables, whereas most of these are still quantitative data, just at the categorical or nominal level. Other researchers use detailed coding of field notes or of unstructured conversations and counts of artifacts as the data for evaluations. Again, in these cases, the observations have been transformed into quantitative variables. Instead, qualitative evaluation is used here to refer to procedures that yield nonnumerical information that helps evaluators to understand the program and the participants' relation to the program, such as full narratives or shorter explanations. To develop this understanding, it helps to learn how participants found the program, what they gained from it, and how they expect to use what they gained in the future. In other words, evaluators want to learn the participants' stories as well as the staff members' stories. Qualitative information helps us to interpret quantitative information and to recognize the unique aspects of different program settings (Corbin & Strauss, 2008).

The first section of this chapter emphasizes data collection approaches particularly central to qualitative evaluation; however, many methodologists (Reichardt & Rallis, 1994) urge that qualitative methods be used together with quantitative evaluation methods, as will be illustrated toward the end of the chapter.

EVALUATION SETTINGS BEST SERVED BY QUALITATIVE EVALUATIONS

The emphasis of this text is on evaluation methods that can be best used with programs having identifiable objectives that can be specified and measured quantitatively in order to learn the degree to which outcomes have been attained. Furthermore, many evaluators believe that with enough care they can discover whether the influence of the program did or did not cause changes in the participants. Given careful participant selection, it is expected that successful programs would be effective if implemented at other locations.

Experienced evaluators often find that they are asked to conduct evaluations when these expectations cannot be met and indeed need not be. In order to develop support for a program, innovators may make their goals intentionally vague (Cook, Leviton, & Shadish, 1985). Vague goals permit different stakeholders to read their own goals in the program's objectives. Following the "letter of the law" of the suggestions in Chapter 5 might lead an evaluator to decline to conduct evaluations of such programs. However, high quality evaluations are still possible in many such settings. Before getting into specific methods, consider some other settings in which a quantitative-focused approach might be applied, but a more probing approach would provide more suggestions to improve the programs.

An After-School Program

Examine the logic model presented in Chapter 2 describing an after-school program for middle school children designed to increase their school achievement, provide skills to avoid some of the negative influences in their communities, and help them develop more respect for peers and adults. An evaluation of such a program could focus on classroom grades, high school graduation rates, and delinquency rates of those who have participated in the program compared to those who did not. However, those children coming to (or being sent to) such a program were probably doing better than other students even before the program started. Thus, a more in-depth examination of what actually happens during program times and learning what the program means to the children and their parents would be more informative. Further, understanding how the high school student leaders and the adult volunteers feel about the program would shed much light on the fidelity of the implementation. The high school students and volunteers could also tell the evaluators what the program means to other people in the community. Because teachers usually learn more than students, it is quite possible that the high school students being paid to lead might gain a great deal from this job.

Evaluating a Political Campaign

Imagine a politician running for a state-level office who is dissatisfied with the voters' reactions to his campaign. Not surprisingly, he would not turn to a professional program evaluator for an experiment. A political campaign is a diffuse, complicated endeavor. Although the ultimate goal of a political campaign is very objective and quantitative, votes cast is not a useful index of campaign effectiveness if one wants to improve the campaign. An approach is needed that can provide recommendations that can be put to use quickly.

The problem faced by this politician is similar to many other situations in which policy must be evaluated before the ultimate outcome index is available. Major, expensive evaluations have been criticized because it took too long for the results to become available; when the evaluations were completed, they were no longer relevant to the information needs of the policy makers. When, for example, an evaluation of a welfare reform plan (called the "negative income tax") was completed, Congress and the White House were no longer in a mood to even consider the innovative plan, regardless of the findings of the evaluation (Cook et al., 1985).

These three hypothetical evaluation programs—a strengthened freshman year program, an after-school program, and a political campaign—are presented to illustrate situations that need evaluation approaches quite unlike monitoring with an information system or using the research designs to be presented in following chapters. Crucial characteristics of the situations calling for qualitative evaluations include (1) a longer cycle between program activities and outcomes than is expected for classroom lessons, medical treatments, or job training; (2) indexes of success that are based on the whole program rather than on measures of individuals such as medical tests or ratings of improvement; (3) a need for results in a short time; (4) multiple stakeholders perhaps with conflicting values; (5) a request for suggestions for improvement rather than just an

evaluation of outcomes; and (6) a concern that the social setting of the program or the program itself is unlike those found elsewhere. Most qualitative evaluations are conducted in settings with at least some of these characteristics.

Although this chapter highlights qualitative methods, issues that require qualitative approaches have been mentioned throughout this text. Almost all work with stakeholders, such as trying to understand what information is most important to them, is part of the qualitative approach to program evaluation. In previous chapters, the importance of verifying that a program has been implemented was stressed. Since it is impossible to know what will go wrong ahead of time, qualitative observations are important in the evaluation of implementation and in the detection of side effects.

GATHERING QUALITATIVE INFORMATION

Before we examine observation and interview techniques, the central importance of the evaluator in making qualitative observations must be stressed.

The Central Importance of the Observer

The single most distinctive aspect of qualitative research is the personal involvement of the evaluator in the process of gathering data. In other forms of evaluation, measurement procedures are designed to dissociate an evaluator from data collection. This is attempted by using written measures of the achievement of goals or objective data such as income level or productivity. The criteria of program success are defined before the evaluation begins and measured with little involvement of the evaluation team except to verify that the data are being gathered as planned with appropriate care. In this way, a form of objectivity is gained. Proponents of qualitative program evaluation, however, would argue that something is lost as well.

What is lost is the evaluator's opportunity to understand the program from the inside rather than standing aloof on the outside (Fetterman & Wandersman, 2004). The opportunity to respond to the information as it is gathered is also lost. In a real sense, the qualitative evaluator is viewed as the measurement instrument. Qualitative evaluators are intimately involved in data collection so that they can react to the observations made. Such reactions may involve adjusting the focus of an interview or even the evaluation itself. For example, it may become evident early in the process of observation that the staff's expectations are incorrect or that the program sponsor's objectives are different from what the program participants expect. An inflexible evaluation plan based on a misconception yields useless information.

Some evaluators object to qualitative evaluations due to fear that evaluations will become very subjective. The loss of the credibility of the evaluation process would be disastrous for evaluators who have striven to demonstrate that evaluations can be used in improving organizations. Qualitative evaluation can be rigorous; however, rigor means something different from what is needed for basic research. Although qualitative evaluators recognize the difficulty, even the impossibility, of finding one correct view of a program, this does not mean that all interpretations are equally legitimate; it is not true that anything goes.

How does one decide when a conclusion based on qualitative information is credible? Consider the tests of the validity of conclusions that are used in daily life. In a mystery story, readers can correctly conclude that "the butler murdered the nanny" by integrating evidence from different sources into a credible, persuasive story. Even when an eyewitness is not available, a defendant can be convicted when caught in a web of evidence. Many court decisions stand on qualitative syntheses of a great number of observations. The tests of validity of court decisions are similar to the tests of validity of qualitative evaluations. Such tests involve corroborating conclusions with evidence from multiple, independent sources; developing a sense of the correctness of the conclusions; and confirming the conclusions with people who know the program.

Observational Methods

Nonparticipant Observers

Since the goal of qualitative evaluation is to understand the program, procedure, or policy being studied, it is essential for the evaluator personally to observe the entity being evaluated. Nonparticipant observation means

that observers are present in the program setting, but they serve no role in the administration or the delivery of the service. Nonparticipant observers would probably not be disruptive of the after-school program described in Chapter 2. They could develop a good understanding of the spirit of the leaders and the enthusiasm of the middle school children.

Qualitative evaluators seek to develop an understanding of how a program operates; they do not seek to answer predefined questions. It is impossible to specify beforehand the details that reveal how the program operates and that end up being important guides to understanding it. If it were possible to list all the things one would be looking for, qualitative evaluation would not be necessary; a checklist could be developed and a clerical assistant could be sent to make the observations. The evaluator must make the observations in order to detect what is important and discover how the details fit into the overall understanding of the program. Avoiding mental blinders and remaining open to many possible interpretations are important contributions of qualitative evaluations.

Using nonparticipant observers to gather information is practical when evaluators can be sure that their presence would not change the social system of the program. But the presence of observers can lead the program staff to act in guarded ways in an effort to control the impressions of an evaluator. Observers can also make staff members nervous and lower their effectiveness, as may occur if staff members had not been informed that evaluators were going to be present. When observations become part of normal operating procedures, staff members become surprisingly relaxed. Nonparticipant observations would be most feasible in settings that are relatively public, such as schools, libraries, and businesses.

Participant Observation

When the services of a program are too private to permit a nonparticipant observer to be present or when the staff members are so defensive that they would not be able to carry out their duties, it may be necessary to use a participant observer. A participant observer has a legitimate role in the program. An assistant in an emergency room would understand emergency room practices. A secretary in a personnel office or a dispatcher in a local police station could provide rich information about the effectiveness and the problems of the service being evaluated.

Using participant observation without the consent of the program staff would be unethical and incompatible with the views of most evaluators. Approaching evaluation without the agreement of the people whose program is being evaluated violates the spirit of mutual trust that is important in effectively functioning agencies. A better approach would be to explain clearly the part the participant observer plays in the evaluation. Cooperation often can be obtained when evaluation is described as a way to learn about and represent the program to those who do not understand how services are delivered and what problems staff members face. Clearly there are less expensive ways to manage a program and one does not need to go to the trouble of conducting an evaluation if service improvement was not the goal of the evaluation. If staff members understand that, cooperation may be obtained.

A variant of participant observation involves a pseudo-participant going through the system. Sometimes the experiences of pseudo-participants are indeed used to gauge the quality of normal treatment. Some stores assign staff (or hire "mystery shoppers") to act as though they are customers in their own and competitors' stores in order to assess the quality of services provided (see Stucker, 2004). People can approach public and private agencies as people would who are carrying out their own affairs.

Examining Traces and Documents

Traces refer to a wide variety of tangible items associated with living. Archeologists examine traces of societies that have disappeared. Running a program also produces many traces. The records mentioned in Chapter 4 are one form of trace, but there are numerous traces produced by any program. In schools, traces include graded homework, teacher lesson plans, tests, litter in hallways, graffiti on walls, student club minutes, student newspapers and yearbooks, athletic trophies, and damage to desks and lockers. When evaluating a community development program, evaluators might count abandoned cars and discarded appliances and furniture in yards or vacant lots, or look for litter in yards and alleys, as well as carefully maintained yards, flower beds, families

in parks, and newly established small businesses. Pictures taken before and after the intervention might be used in evaluations.

Physical traces add a dimension to an evaluation that is hard to gain in other ways. Someone conducting an evaluation of a school program would gain considerable understanding of the school and community by systematically examining the traces suggested above. This understanding could not be gained without the personal presence of members of the evaluation team. Although evaluators may well consider the traces that might be observed, qualitative evaluators do not begin making observations with a prepared list of items to look for; whatever is important will be discovered through the observations.

The meaning of any one physical trace is nearly impossible to understand. It is the accumulation of evidence from many traces that leads to a conclusion. Furthermore, tentative interpretations are compared with information obtained through interviews and observations of behaviors. By seeking various forms of information from different sources, evaluators seek to draw conclusions that are collaborated in several ways. Qualitative evaluators borrowed the idea of triangulation from surveyors to describe this process. By triangulation qualitative evaluators mean that they use different approaches to learning about the program.

Documents refer to official and unofficial statements made by program sponsors, planners, board members, etc. Qualitative evaluators often learn how the program is intended to function in documents. Records kept by the staff and program managers are also documents of interest.

Interviewing to Obtain Qualitative Information

Qualitative interviewing is different from simply administering a written survey orally. A structured survey given orally remains a structured survey. Qualitative evaluators use open-ended, unstructured questions in order to learn detailed information about programs. Interviewers want the respondents to use their own words, thought patterns, and values when answering the questions (Patton, 1980).

Another method for gathering qualitative data—focus groups—has already been addressed in Chapter 6, so that information will not be repeated here. Still, focus groups may be a good option when qualitative data is desired. Some evidence suggests that in-depth interviews provide the most cost-effective data collection (Namey, Guest, McKenna, & Chen, 2016), but many evaluators find that focus groups provide important benefits. Others suggest that the method of dyadic interviews combines some of the advantages of interviewing individuals with some of the benefits of focus groups (Morgan, Eliot, Lowe, & Gorman, 2016).

Preparing for the Interview

Before conducting interviews, evaluators make sure that the interviewee understands the purpose of the interview and has consented to be interviewed. There is no point in trying to mislead the individual as to the purpose of the interview. Such unethical behavior may well create controversy and distract from the credibility of the evaluation report. It is a good idea to confirm the appointment for an interview a day or two beforehand. Interviewers should be on time, even early. One will have better rapport if interviewers dress according to the norms of the organization. One need not dress in a designer suit to interview a business executive, but do not show up wearing a baseball cap.

Developing Rapport

Interviewees usually are guarded as interviews begin. A personal relationship is needed to establish rapport and trust between the interviewer and respondent. Rapport can be fostered by asking some orientation questions and showing accepting, friendly reactions to the answers. Starting out asking, "How did you become a librarian?" gives a librarian a chance to talk and relax.

Asking Questions

One of the most crucial points in qualitative interviews is to avoid using questions that can be answered "Yes" or "No." Patton (1980) provided several sets of questions that could produce useful information only if the

interviewee spontaneously expanded upon answers to the actual questions. Figure 8.1 was inspired by Patton's suggestions. As the questions in the figure show, qualitative interviewers use open-ended questions that encourage the respondent to talk and elaborate. The best questions begin with phrases such as "What is it like when … ?" "How would you describe the employees at … ?" "Please tell me how … ?" or "What happens if a client complains about … ?" Someone can refuse to answer questions completely or may provide misleading answers; however, the format of such questions encourages the informant to provide answers revealing important information.

Questions Not Likely to Elicit Useful Information from an Interviewee	Open-Ended Interview Questions That Will Encourage the Interviewee to Provide Information
Is this a social psychology graduate program?	What is the name of this graduate program?
Are you the director?	What is your role in this program?
Is the emphasis on applications of social psychology?	Please characterize the emphasis of the program.
Do the students enter the program with interests in applications?	What are the interests of the applicants to the program?
Do the courses have an applied orientation?	How do the courses relate to the emphasis of the program?
Do the students like the applied orientation?	How do the students react to the applied orientation?
Are the theses and dissertations related to the theme?	How do the theses and dissertations fit into the program's theme?
Do the students seek teaching positions after graduation?	What type of positions do the graduates seek?
Do the graduates get good positions?	How do graduates use the skills learned from the program?

Figure 8.1 Two possible sets of questions for a director of an applied social psychology graduate program. The contrasts between the two columns are the important points of this figure. Do note, however, that the qualitative interviewer would not have an inflexible set of questions ready to use. One would have the issues clearly in mind, but the actual questions asked would develop as the interview progressed.

Probing for More Information

Sometimes an interviewer senses that there is something more to be said, or the interviewee has not understood the point of the question. Conducting qualitative interviews gives evaluators the opportunity to tailor questions to the respondent. The purpose of probes is to gain more information. One behavior that will encourage respondents to answer more fully is easy: interviewers often nod their heads as the respondents speak. This sign of encouragement can often be seen during televised interviews as a journalist keeps nodding while a newsmaker talks. An interviewer is seeking information; consequently, it is important that the interviewer treat everyone in a nonjudgmental manner. One nods to encourage the speaker whether the interviewer agrees or disagrees with what is being said. A generic probe is silence. Instead of bringing up a new question when the respondent has stopped talking, the interviewer can simply remain quiet while looking at the respondent: being quiet for five seconds has a surprisingly strong impact.

Clearly either approach can be overdone and become artificial. More active probes are also used. The interviewer can probe by simply asking, "Can you tell me more about … ?" Interviewers may want to check on an interpretation of what was said by asking, "Let's see, if I understand you, you said that …. If that is correct, can you tell me what happens next?" It is critical to avoid using directive or leading probes. An interviewer who interprets a tone of voice as anger would be using a directive probe if the follow-up question is, "Why does that make you angry?" A nondirective probe would be, "How did you feel when you answered that last question?" In this way the interviewer has responded to the appearance of emotion and can learn about it without running the risk of labeling it incorrectly.

Recording Answers

Although there was a time when the logistics of recording conversations were cumbersome and made too many people uncomfortable (Lincoln & Guba, 1985), nearly unobtrusive equipment and the pervasive presence of other electronic devices like smartphones make it an excellent option now. Further, computer programs are available to organize the verbal material (Weitzman & Miles, 1995) and voice recognition software to convert spoken language into text files has become readily accessible (e.g., Dragon Naturally Speaking, 2018).

Ending the Qualitative Interview

A structured interview ends with the last prepared question. By contrast, a qualitative interview could go on and on. When the scheduled appointment is nearly over, when the participants are fatigued, or when the information being discussed has become redundant, it is time to quit. It is best for interviewers to summarize the major points that have been made by saying, for example, "Here is how I understand your views. You have said that you enjoyed teaching the special freshman course, but that you received very little information from the department chair or the dean about how these courses were supposed to differ from regular sections of the same course. You said that you found that frustrating—is that correct?" The advantages of summarizing include getting a chance to verify interpretations before leaving and permitting respondents to expand on some point. Finally, interviewers thank respondents for their attention and thoughtful answers. One might ask if it would be possible to check back later for clarification if necessary.

CARRYING OUT QUALITATIVE EVALUATIONS

Although naturalistic evaluations require more involvement and creativity on the part of the evaluator than do traditional evaluation methods using surveys and checklists, the essential plan of a qualitative evaluation is quite similar to how people go about learning anything. The steps to conducting a qualitative evaluation are given in discrete phases below; however, readers should not view these phases as a step-by-step recipe. The phases overlap and later phases provide feedback that could lead an evaluator to revise initial conclusions. Nevertheless, at different times during an evaluation, there are types of activities that are more important than at other times.

Phase I: Making Unrestricted Observations

Qualitative evaluations begin with observations of the most crucial program events, activities, written materials, and settings. Qualitative evaluators do not seek to observe a random selection of activities or people. Often they seek out those who know more than others and can provide more information on how things work in that setting. At times, deliberately selecting extreme examples, such as some who demonstrate the most of some characteristic or behavior and others who show the least, may help accentuate relevant factors. Although there are no restrictions on the observations and interviews, evaluators try to direct the information-gathering process toward important elements of the program. Observations, impressions, and interviews are recorded in field notes. The chances of drawing erroneous conclusions increase when qualitative evaluations are based on samples of settings, people, or conditions that happen to be convenient or that are determined by program managers. Qualitative evaluators should have unrestricted access and should be permitted to gather information from all aspects of the program or agency.

Phase II: Integrating Impressions

The second phase, which actually begins with the first observations or interviews, is to integrate the impressions that are formed during this unrestricted observation period. From these impressions evaluators develop some specific ideas about how the program is run, how services are provided, how participants understand the purpose of the agency, and so forth. Then further observations and interviews are conducted in order to check on the accuracy of these impressions and to seek information on topics that were not explored

thoroughly. With this additional qualitative information, evaluators refine their initial impressions. When additional observations no longer change the impressions, the major part of the information-gathering phase is completed. The field notes are to be saved so that others may see them. For a list of sources for online guidance and other assistance, along with active links, see the eResources.

Phase III: Sharing Interpretations

As impressions are formed, qualitative evaluators share their views with stakeholders and other evaluators. The qualitative approach has been misunderstood by some critics as subjectivity running wild. Describing tentative impressions to others who know the program and have a stake in it is a way to obtain feedback so that impressions do not reflect the prejudgments of an evaluator. People familiar with the program can provide additional information to correct misunderstandings. Experienced, but uninvolved, evaluators can challenge interpretations that are not adequately supported.

Phase IV: Preparing Reports

Once checks with stakeholders and colleagues verify the accuracy of the impressions that have been formed, an evaluator is able to present the descriptions of the program and to draw evaluative conclusions about the program. Reports are usually lengthy. One of the central goals of qualitative evaluations is to provide detailed descriptions of programs through the eyes of the stakeholders along with the insights of the evaluator. Reports thus typically include substantial chunks of what has been said verbatim. The job of the evaluator is then to integrate the views of many stakeholders so that everyone understands the program better than before. Although the reports contain details to communicate an understanding of the program, stakeholders who need information promptly can be given shorter, more focused reports.

The report is not presented as a definitive evaluation of the program since additional information can come to light later, conditions may change, and the membership of the stakeholder groups may change. The findings of qualitative evaluations can be applied in other locations to the extent that other settings are similar to the one evaluated. Since qualitative evaluators are sensitive to the many specific factors that can affect program success, generalizing findings to other settings is done only with extreme care.

Are Qualitative Evaluations Subjective?

It would not be surprising if many readers reacted to this presentation with skepticism; it just seems very subjective. Two comments are offered in response. First, some years ago the second author was discussing research methods with an experienced social psychologist. The conversation concerned the focus of this chapter—traditional quantitative research procedures versus qualitative personal observations. When asked, "If you wanted to learn about a group you knew nothing about, would you go there and live among them or would you send surveys and questionnaires?" the more senior professor responded, "Go and live there." Although we both would have preferred direct involvement as the mode for personal learning, at that time neither of us was bold enough to supplement the quantitative techniques we used with qualitative observations. Qualitative methods permit the development of understandings that cannot be obtained with predefined surveys and checklists. When evaluators understand programs very thoroughly, their evaluations can contain reasonable recommendations that are likely to be used.

The second response to the concern about the subjectivity of qualitative evaluations centers on the question of how objective other evaluation approaches are. Subjectivity is not observed in the scoring of surveys or in the analyses of program records. However, qualitative evaluators are quick to argue that the choice of variables to measure and decisions about what is important to control and what does not need to be controlled can also affect the findings of evaluations (Guba & Lincoln, 1989). Although the ways prejudgments can affect evaluations are different for quantitative and qualitative evaluations, evaluators must be equally wary of unexamined assumptions regardless of the form of evaluation (Reichardt & Rallis, 1994). Let the methodologist who has no shortcomings cast the first stone.

COORDINATING QUALITATIVE AND QUANTITATIVE METHODS

For most evaluators, the question is not whether to use qualitative methods *or* quantitative methods; the question is how to use the methods so that they complement each other in the best ways possible (see Campbell, 1987; Maxwell, 1985; Reichardt & Rallis, 1994; Rossman & Wilson, 1985; Silverman, Ricci, & Gunter, 1990). Depending on the evaluation questions to be answered, the relative emphasis on each approach varies. Light and Pillemer (1984, p. 143) write, "The pursuit of good science should transcend personal preferences for numbers or narrative."

Bringing Numbers to Life

Although a political campaign requires primarily a qualitative evaluation and rapid feedback, the candidate would probably also like a careful accounting of the allotment of campaign funds and a quantitative analysis of opinion polls. The effectiveness of an experimental medication requires primarily quantitative measures indicating improved health and long-term follow-up of patients' mortality and an enumeration of side effects. Medical research is improved when the focus on length of life is complemented with an examination of the quality of life (McSweeny & Creer, 1995). Suppose that the after-school program mentioned as this chapter began did show an increase in rate of high school graduation for participants. All stakeholders would be pleased, but to supplement that finding with some stories of how the program helped a student, even a story from a parent of a participant, would make that finding memorable and compelling in a way that the difference between two percentages cannot.

Getting Insights from the Most Successful Participants

The success case method was mentioned in Chapter 2. If the participants in a job training program wanted to learn how to make the best use of the program, they might learn from those who have completed the program and obtained a stable job. If the staff of such a program wanted insights in how to improve the program, they too might gain some ideas from those who have done well. Of course, it is true that some participants do well after a program for reasons that have nothing to do with their experiences in the program. A qualitative evaluator looking for successful participants would want to note that some participants did not really need the program. But among those who appear to be typical participants when they started, there are some who have done particularly well. Those former participants might well help staff to learn which activities are less helpful and which are more helpful. They can also tell the staff how they used the skills they were learning. How did they go beyond the activities defining the program? Although the staff has resources that the participants do not have, staff members cannot completely put themselves into the shoes of the participants. Only a qualitative evaluation approach can aspire to gaining insights using such an approach because the critical variables and insights cannot be anticipated and, therefore, cannot be discovered using a survey (see Brinkerhoff, 2003). Note that the strength of such an approach does not focus on the outcomes of the typical, or average, participant as so many evaluation methods do. This is not to say that average results are not important; they are. But creative ideas are seldom suggested by a focus on average outcomes. And although it is not always necessary, sometimes getting insights from the least successful participants may also provide important information.

Gaining Insights from What Works Well

The qualitative evaluation method called appreciative inquiry (AI) was adopted from organizational development (Preskill & Coghlan, 2003). One might begin by asking, for example, "What are you especially pleased with in the after-school program?" The assumption on which AI is based is that in every organization something works well. The best approach to improving an organization is to focus on what works well; a critical job for the evaluator is to learn what works well and to help stakeholders recognize that and build upon it. Further stakeholders are asked what they would like to see in the future, what would be ideal. Other approaches seek to learn how the organization functions. This can be achieved by sampling stakeholders and seeking those who are well acquainted with the program; by contrast, AI evaluators seek to include as many stakeholders of the program as possible because for the program to develop positively, many people must catch the vision of possible futures that they indeed build. AI proponents fault many other approaches to evaluation

as focused too much on problems, thus contributing to demoralizing stakeholders. More will be said on this topic later in Chapters 13 and 14 on Evaluation Reports and Encouraging Utilization.

Changing Emphases as Understanding Expands

An evaluation planned initially as a quantitative evaluation can be enriched by adding a qualitative component when unexpected, negative side effects are discovered. Often evaluators are slow to notice side effects because unexpected effects, good or bad, cannot be anticipated. Evaluation procedures also change when observations cannot be made as planned. In order to make any interpretations of the data it may be necessary to utilize a more qualitative approach.

In a similar fashion, a qualitative evaluation can become more quantitative as the evaluation questions become focused and clear. Once qualitative impressions are well developed, it might be possible to form hypotheses about the expected pattern of some quantitative measures of program outcome. In this way, qualitative evaluators can strengthen some aspects of their conclusions.

The Evaluation Questions

Evaluations have been classified as evaluations of need, process, outcome, and efficiency. Certain questions are more likely to require an emphasis on qualitative methods, whereas others will require quantitative methods. It is not possible to compare program costs with outcomes when using only qualitative information. However, most evaluations have multiple purposes. Implementation must occur before one can expect outcomes to be observed. Evaluation of implementation can be quantitative, such as determining the number of sessions, the number of participants, and the proportion of participants that met the official eligibility criteria. But it is also important to know how healthy the interactions between staff and participants are, and how participant families make the decision to participate. It is not unknown for intra-staff conflicts to develop or for staff members to defend what they consider to be their turf and thereby damage the program. In the areas of social and health services, such issues are very important in learning how to improve programs. Since so many evaluation questions are complex, it seems quite reasonable to expect that those involving both qualitative and quantitative approaches would be better than evaluations limited to one or the other (Datta, 1994).

Cost of Evaluation

One concern with qualitative methods is that the cost, especially for an evaluation of a major program with sites throughout the country, would be too great. Sending observers to the sites or recruiting and training individuals in many locations would be difficult and indeed quite expensive compared to sending surveys to and obtaining records from the sites. The expense is compounded by the open nature of qualitative evaluation. Someone commissioning a qualitative evaluation must trust that the work will produce useful information even before the evaluators know what observations will be made or how the observations will be used in reports (Lincoln & Guba, 1985). One strategy to contain costs is to combine qualitative methods with quantitative methods and use the two methods together to verify particular issues such as an appropriate assessment of needs or the credible implementation of the program. In addition, some approaches to qualitative analysis that take advantage of newer technology reduce the costs of time and money substantially (Neal, Neal, VanDyke, & Kornbluh, 2015). With smaller programs, the distinct advantage of richer data may be much more valuable than a slight increase in cost.

PHILOSOPHICAL ASSUMPTIONS

Some proponents of qualitative evaluation methods view the use of qualitative research procedures as flowing from a view of knowledge that is distinctly different from that held by evaluators who focus on quantitative approaches (Lincoln, 1990). At the most extreme, these writers adopt a constructivist philosophy that considers reality to be only a function of social agreements rather than something outside of people that can be studied scientifically (see Hedrick, 1994). Guba and Lincoln (1989, p. 13) asserted that "there is no reality except that created by people." On the other hand, they have also said, " … there are external constraints that

limit what can be agreed upon" (Lincoln & Guba, 1985, p. 85). Many observers feel that in practice even the most partisan supporters of qualitative-only evaluations don't limit themselves in ways that one might expect on the basis of the subjectivist philosophy they affirm in their more abstract writings (see Datta, 1994).

There is a second theme subject to less confusion that sets the most avid qualitative evaluators apart from the mainstream evaluation community. For most evaluators the purpose of evaluation is to improve the delivery of services to people in need by studying the implementation and impact of programs. Lincoln and Guba (1985) assert that quantitative evaluations are planned and managed in ways that bar some stakeholders from participating in the evaluation. Instead, they argue that evaluations should be consciously directed toward empowering the most disadvantaged groups in society. They call this "authentic evaluation." A number of observers imply that empowering all stakeholders is not possible in many evaluations (Greene, 1987; Mark & Shotland, 1987). There is an important tension between doing better to empower as many stakeholders as possible and recognizing the many realities that limit such empowerment. It may be that Lincoln and Guba are urging evaluators to do more than they are commissioned to do and indeed can do.

As demonstrated by the evaluators whom Reichardt and Rallis (1994) recruited for their edited volume, there is a long and productive record of qualitative methods being used along with quantitative methods by evaluators who do not accept the philosophical assumptions held by some extreme advocates of qualitative evaluation methodologies (see Denzin & Lincoln, 2005; Guba & Lincoln, 1989). Shadish (1995) has shown that the philosophical underpinnings sometimes presented for qualitative evaluation methods are "at best a distraction, and are at worst downright harmful" to efforts to resolve the debate over how to use qualitative methods in program evaluation.

CASE STUDY 8

Healing a Community

Many program evaluations are initiated through formal processes such as a call for proposals or a request directly presented to a program evaluation organization. But sometimes the connection happens more spontaneously, as in this case. As described in Evaluator Profile 8, neither Dr. Steven Kniffley nor the police chief of a nearby city expected their attendance at a panel presentation on the relationship between law enforcement and communities of color to lead to a program evaluation—but it did. In addition to illustrating how qualitative methods can be the appropriate approach in an evaluation, this case also demonstrates how good evaluators must be flexible as they learn about programs, so they can respond appropriately and helpfully to the situation. If anything, these points are especially relevant for evaluators early in their careers.

Meeting Needs

Recent conflicts between communities of color and law enforcement (in response to the deaths of young Black men) highlight the need to explore the assessment of cultural competency in law enforcement departments. Police departments face significant challenges as they adapt to engaging with an increasingly diverse community that has become more competent with higher expectations. These challenges are compounded by the homogeneity of most police departments (e.g., white and male) and a "glass ceiling" effect in the representation of persons of color in leadership. Furthermore, faced with limited numbers to engage in policing growing populations, the perceived least culturally competent group (e.g., younger officers via age and experience) are perhaps the most likely to engage communities of color during high crime times (e.g., evenings), which possibly contributes to more conflict.

Implementation

The focus of the request was to evaluate the department's relationship with the community they served—specifically to assess the community's perception of the officers and potential barriers to establishing an effective relationship that would empower both community members and law enforcement to collaborate on creating safer neighborhoods as well as to determine the officers' reflective experience of multiculturalism, readiness for cultural contact, and the presence of prejudice towards the communities they served. This broad area, approached in an exploratory manner, makes it harder to separate elements of implementation, outcomes, and side effects from each other than in most cases. To adapt the details of this case to the MISSION framework, the elements examined are described in this implementation section, the method Dr. Kniffley used is addressed in the side effects section, and the findings are noted in the outcomes section.

Stakeholders

Obviously the police are stakeholders, as, implicitly, are local government officials. In this case, the whole community should be considered stakeholders, although some members of the community were much more aware of the details of the situation being evaluated. This case illustrates some inherent challenges to connecting with the full range of stakeholders in some evaluations, as well as a creative and productive approach to that challenge.

Side Effects

As noted, this section describes the qualitative data collection process. Among other things, although the exploratory nature of the evaluation did not label particular variables as "Outcomes" and "Side Effects," the method deliberately sought information that would not have been listed as expected outcomes in a formal research design.

Steven conducted three focus groups with police officers in the department and community members. Two of the focus groups were with the police officers and one group was with neighborhood association presidents and community leaders. The police focus groups consisted of six participants each. The focus group with officers engaged them in a structured dialogue consisting of seven questions that were designed to elicit narratives around the following topics: (1) city culture, (2) multicultural issues that impact their jobs, (3) perception of skill and comfort engaging with culturally diverse individuals, and (4) their assessment of the cultural climate of the department. The community member focus group consisted of 12 participants and was structured around nine questions with a similar design to elicit narratives around the following topics: (1) city culture, (2) understanding of police policy and procedures, (3) cultural competency of officers, and (4) diversity training needs for police officers. Dr. Kniffley and a team of his graduate students transcribed and coded the recorded narratives. Using a common code book, individuals on the team coded the narratives for themes and then met together and reconciled any differences.

Improvement Focus

The findings and recommendations in particular should make it clear that Dr. Kniffley approached this evaluation as a way to help the department improve in the future, rather than as a way to criticize how they were not yet perfect. Although qualitative methods are not only useful in formative evaluations like this one, the specific methods of focus groups that evoked stories from important stakeholders were particularly appropriate for this improvement-focused evaluation.

Outcomes

For the police officer focus group, a number of themes emerged from their responses. First, the officers noted significant generational differences with the department with conflict being experienced between older and younger officers. Secondly, the officers highlighted the chronic disparities between certain sections of the city that impact their perceptions of criminal activity. Additionally, several of the officers discussed the role of parents and their influence on the issues/conflict between community members and police officers. Furthermore, the officers discussed their frustrations concerning the community belief that law enforcement's only role is to arrest and the lack of awareness of the multiple roles officers have to fulfill in responding to emergency/nonemergency situations.

For the community members focus group five themes emerged from the analysis. First, the community members noted that differential perception of the cultural challenges facing the city existed between White and Black community members with segregation being seen as a significant issue. Community members also mentioned issues that they perceived to be impacting community/police relationships such as lack of jobs, skepticism towards officers and their motives, and the need to build better relationships with both children and adults.

A number of tangible recommendations based on the results were provided to the chief such as (1) increasing transparency through quarterly community conversations, and increasing access to the complaint process, (2) improving community/police relationships through the development of a community/law enforcement task force, increased law enforcement/youth programming initiatives, and a monthly officer spotlight, and (3) increasing access to diversity training opportunities such introductory Spanish.

Nuances

Although a case could be made that the qualitative approach gathered information about some potential mechanisms, the exploratory, formative nature of this evaluation really deferred such analysis to future follow-ups.

Summary

Not all evaluations take on such difficult, even potentially explosive, issues like this one. Effective evaluations in such settings hold not only the challenge of managing real tensions, but also the hope of making a valuable difference where it is most needed.

Source: S. Kniffley (personal communication, January 31, 2018).

EVALUATOR PROFILE 8

Steven Kniffley, PsyD

Dr. Steven Kniffley is an early career psychologist who, upon completing his doctoral training, assumed that he would go directly into clinical work with underserved children, adolescents, and their families. He had taken a program evaluation course as part of his graduate program (taught by the first author of this textbook), and while intrigued, thought it would not be a significant part of his career. However, he knew that in order to effectively impact the populations he was passionate about (e.g., communities of color and Black males) he had to keep his options open.

As noted in Case Study 8, Steven's first experience with program evaluation occurred by accident. He had been asked to serve on a panel concerning the relationship between law enforcement and communities of color. As the psychologist on the panel, his task was to discuss the psychological barriers that may be impacting the police/community relationship. However, one member of the audience, a police chief from a midsize Midwestern city, saw the potential for this discussion to translate into an assessment of the relationship

between his department and the community they served.

The chief wanted to know how his department was doing in terms of cultural competence, to gain information about the community's perception of the department and their engagement with culturally different individuals, and to consider any recommendations to improve the community/police relationship. Although nervous, Dr. Kniffley took this opening and ran with it, developing a thoughtful and informative program evaluation that incorporated voices from multiple stakeholders (e.g., officers, community members) and utilized a number of methods (e.g., surveys as well as focus groups) to collect relevant data. In the end, Steven was able to provide a comprehensive report with tangible and implementable recommendations that led to a number of important changes in the relationship between the law enforcement department and communities of color over the next several years. With this experience, Dr. Kniffley was initiated into the world of professional program evaluation.

Since his first program evaluation, he has gone on to serve as an evaluator for a number organizations such as hospitals, nonprofits, city departments, and school districts. His evaluation experience has also taken him to other countries such as South Africa where he evaluated a U.S.-based leadership training program for South African high school students. He notes two particularly important priorities of his process of conducting program evaluations: (1) taking the time to include as many voices and stakeholders as possible into creating the organization's narrative, and (2) crafting the final evaluation product in a way that is understandable to multiple audiences. He believes that no program evaluation is meaningful unless the stakeholders for whom it is produced can read it and find that the recommendations indicated are relevant, resource appropriate, and respectful of the differences present in the community being served.

Source: S. Kniffley (personal communication, January 25, 2018).

SUMMARY AND PREVIEW

There are settings in which evaluators cannot use the research methods based on the social science research methods model. Some complex programs requiring an evaluation in a fairly short time cannot be approached using only quantitative methods. Direct observations are usually required to understand programs fully. Furthermore, recommendations for program improvement are most likely to be applicable when evaluators have a good understanding of programs and their stakeholders, especially the successful participants. In addition, evaluators who demonstrate a deep understanding of the strengths of programs find their recommendations viewed as more credible than reports based on numbers alone. Qualitative methods (talking, watching, and drawing conclusions) appear simple, but are quite complicated as evaluators seek to integrate the information they gather. Nevertheless, the use of qualitative methods in conjunction with quantitative methods strengthens program evaluations.

When the objectives of programs can be specified and when an impact model has been developed, one can use the traditional research methods that have been developed in the social sciences. These forms of program evaluation methodology become the focus of the next chapters.

STUDY QUESTIONS

1 Most college course evaluation forms are largely made up of survey items with five- or six-point rating scales. Here are some items taken from one form:

   a The goals of the course were clearly expressed at the beginning of the term.
   b The course helped me to develop intellectual skills, such as critical thinking or problem solving.

   Rephrase these into open-ended questions that a qualitative evaluator might use in an *interview* to learn about these same issues. Then suggest some responses to these open-ended questions that could not be expressed by a numerical rating.

2 Look back to Chapter 7. Develop some suggestions of how qualitative methods could supplement and clarify the meaning of the summaries of the quantitative material the evaluators gathered. What interview issues could a qualitative evaluator address that would be very difficult to handle using a written survey? What are some questions that appreciative inquiry might suggest?

3 Many observers worry that internal evaluators are under pressure—sometimes overt, sometimes subtle—to produce favorable evaluations. Discuss how such pressures might affect qualitative evaluators.

4 Think back to the program described on the first page of this book. How would an evaluator use qualitative evaluation techniques to gather information on the effectiveness of a sexual assault prevention program offered by a college counseling center?

5 Suppose that a six-month parent education program was developed at an urban university and successfully implemented with mothers who gave birth at a public hospital. When the same program was offered to families qualifying for public aid in several rural areas, young mothers seldom returned for even

the second or third session. Use the ideas in this chapter to develop some approaches to learn why the program was valued in one setting, but thoroughly rejected in the second setting.

ADDITIONAL RESOURCE

Patton, M. Q. (2003). Qualitative evaluation checklist. Evaluation checklists project. Western Michigan University Evaluation Center. Retrieved from http://www.wmich.edu/sites/default/files/attachments/u350/2014/qualitativeevalchecklist.pdf.

Patton provides a thorough overview of qualitative evaluation methods and procedures. This is a checklist, so points are simply made without explanation; however, it is a remarkable guide uncluttered with philosophical jargon. As noted in the eResources, the Western Michigan University Evaluation Center—http://www.wmich.edu/evaluation—with specifics such as their Evaluation Checklist page —http://www.wmich.edu/evaluation/checklists—provides substantial resources for evaluators at all levels.

9 Outcome Evaluations with One Group

Third grade, a class for diabetic senior citizens, and a parenting skills workshop are programs that begin and end at specific times, are directed to a group of people with similar needs, and are planned to impart similar benefits to all participants. The most common approach to evaluation in such settings is to examine how well participants perform after completion of the program. This chapter presents the value as well as the weaknesses of such simple evaluations. Then, the following chapters illustrate how these basic approaches are augmented to improve the validity of interpretations. The reader should develop a sense of how the introduction of each level of control increases the power of the evaluation to demonstrate that experiences in the program helped the program participants to achieve the program's objectives. The methods in this and the next two chapters are more similar to common research designs than approaches to evaluation discussed so far.

ONE-GROUP EVALUATION DESIGNS

There are two basic options with one-group designs—to assess various elements only after the program or to assess elements both before and after the program.

Observe Only after the Program

When teachers, nurses, administrators, or judges inquire into the success of programs, they frequently use the simplest form of outcome evaluation because they want to know how the participants are faring after a service has been provided. Do members of a job-training group have jobs three months after receiving job skills training? How many of the people who attended a smoking cessation clinic are in fact not smoking one month after completing the program? The first step in deciding whether a program is useful is to learn how many of the participants finish the program with a level of achievement that matches the program's goals. All that is needed is a set of systematic observations of participants at some specified time after completion of a program. Note, however, that such an approach to evaluation does not even show that the participants changed during the time they were in the program.

Observe before and after the Program

The pretest–posttest design is used when stakeholders want to know whether the participants improved while being served by a program. For example, students should read better after a year of reading instruction, and cholesterol levels should be lower after patients have begun taking a statin pill each day. The program might have caused the improvement; however, information gathered from such a design does not permit such an interpretation because there are many alternative interpretations that remain viable even when a comparison between a posttest and a pretest reveals that change occurred. These alternative interpretations have been called "threats to internal validity" by Campbell and Stanley (1963); these threats will be discussed later in this chapter.

This text is based on the assumption that an evaluation should be designed around the questions to be answered and the complexity of the program. Evaluation is not basic research; evaluators do not assume that an evaluation must be designed to answer the same type of questions that basic research studies would address. When programs are relatively inexpensive, not harmful to participants, and fairly standard, rigorous evaluations are not always needed (Smith, 1989). It is important, however, for evaluators to recognize when a more ambitious design is appropriate.

In some situations the evaluation plan cannot be any more complex than either of these two designs. The Government Performance and Results Act (GRPA) of 1993 was passed to shift the focus of government decision making and accountability away from simply listing the activities that are undertaken by a government agency—such as grants made or inspectors employed—to the results of those activities, such as gains in employability, safety, responsiveness, or program quality. Under the Act, agencies are to develop multiyear plans, analyze the extent of the achievement of the goals, and report annually. Although GRPA was initially a requirement for federal agencies, many states in the United States have adopted similar laws. The results of the activities of government agencies are evaluated by examining what was achieved, but these results can only be found on one "group." There is no possibility of, for example, comparing scientific activity related to the National Science Foundation (NSF) with the level of activity in the absence of the NSF. Similarly there is no possibility that Social Security payments would be suspended for a year to measure the level of poverty among the elderly without Social Security. The point is that the only way such national programs can be evaluated is through an examination of key variables after programs have been implemented. As you think about the limitations of such approaches to evaluation, bear in mind that the evaluations of many large-scale, costly government programs are, at best, posttest-only or pretest–posttest designs.

USES OF ONE-GROUP, DESCRIPTIVE DESIGNS

Did the Participants Meet a Criterion?

In many job-training settings there are objectively defined skills that trainees must attain in order to do a specific job successfully; in such settings a posttest-only evaluation is satisfactory. A number of writers suggest that a criterion of successful psychological counseling is the point when measures of mental health cannot distinguish between treated people and people without diagnosed emotional problems (Jacobsen & Truax, 1991). Compliance with a law can be measured using a posttest-only design. Manufacturing firms establish indicators of quality called benchmarks to track how well their firms are doing compared to firms thought to be the best performers in their industry. If there is a way to set a criterion of good performance and if performance can be measured validly and reliably, then comparing participants' outcomes with benchmarks may be a perfectly satisfactory evaluation design.

Did the Participants Improve?

An important concern is whether the participants changed in the direction that the program was planned to encourage. Do rehabilitation patients speak and walk better after taking part in a rehabilitation program? Case Study 9 describes how a pretest–posttest design was used to evaluate a small program to address spiritual needs among homeless adults. Certainly, improvement is a fundamental objective of programs; however, discovering that change occurred during a program should not be confused with believing that the same change would not have occurred without the program. Some stakeholders appear to believe that a statistically significant difference between a pretest and a posttest provides evidence that the program caused the improvement. A statistically significant finding can only show that the change was unlikely to reflect only sampling error, not to reveal causality.

Did the Participants Improve Enough?

If an improvement from pretest to posttest is statistically significant, evaluators face another issue: did participants improve *enough* to demonstrate a real effect in their daily lives? If the outcome of a school program was to change attitudes on about civic responsibility, an increase of 3 points on a 92-point attitude test does not seem to be much of a change even if statistically significant. It was probably not substantively significant. Clinical psychologists also advocate research on assessing meaningful change for clients (Meier, 2008).

There are a number of ways to assess change; some way to decide whether the change is enough must be found. No approach is definitive; all require thoughtful reflection. Two approaches will be discussed: (1) when an outcome variable itself is meaningful (e.g., successfully holding a job) and (2) when the outcome variable

serves as a proxy for a variable that is too difficult or expensive to measure (e.g., improved attitudes about holding a job and improved skill in filling out a job application form rather than actual success on a job).

Outcomes That Have Meaning in Themselves

Imagine a smoking cessation program. Since people buy cigarettes in 20-cigarette packs, smokers can report fairly accurately how much they smoke per day. And they can report if they have quit. Quitting would yield the most benefit, but reductions in the rate of smoking are also valuable. A meaningful index of the outcome of a smoking cessation program would be the difference between the number of cigarettes smoked per day before the program and the number after the program.

$$\text{Reduction} = \text{Number smoked}_{\text{preprogram}} - \text{Number smoked}_{\text{postprogram}}$$

where *Number smoked* is the number of cigarettes smoked per day. After such an index is calculated for each participant, the mean among all participants could be found. A second index would be the percentage of participants who are no longer smoking after the program. Comparing the reductions occurring during a smoking cessation program with the reductions of other similar programs would permit one to learn if a program being evaluated produced a substantively significant outcome. And it is probably obvious that a reduction of a full pack a day on average is a substantial one, whereas a decrease from three packs plus one cigarette a day to three packs is not.

Fitness programs would also permit reporting program outcomes in meaningful units. Increases in maximum weight one can lift, reductions in time needed to complete a lap around a track, and lowered heart rate while doing some specified activity are three meaningful indexes of the results of a fitness program. To decide whether the change was large enough to be of value requires comparisons to norms for heart rate appropriate to the age of the participant.

Outcomes That Are Measured Using Proxy Variables

It is often the case that evaluators cannot measure what really is the desired outcome from program participation. College faculty want to train educated people, but we settle on calculating a grade point average or using a score on a standard test. Counseling programs are planned to encourage participants to develop into well-adjusted people, but we settle on learning whether former participants score lower on tests of anxiety, stress, or depression. Such variables serve as proxies for the real thing. Because such variables could be measured with a variety of tests that vary in the number of items that they contain, neither a single score nor the mean score for a group of participants can tell us anything precisely about the value of the program. We can, however, standardize the difference between the preprogram and the postprogram means to find a standardized effect size. If the program is designed to reduce an undesirable behavior or emotion, the standard formula for effect size is as follows:

$$g = \frac{M_{\text{preprogram}} - M_{\text{postprogram}}}{S}$$

In words, the difference between the mean prior to participating in the program and the mean after the program divided by the standard deviation of the variable measured (e.g., anxiety) is the effect size in Cohen's d units (see Posavac, 1998; Zechmeister & Posavac, 2003). Once standardized, we can compare the effect size of the program being evaluated to the results of other programs, and there are some general guidelines to interpreting effects sizes on their own. Here is an example. A counseling program was designed to reduce test anxiety among first-year students; assume these means:

$$M_{\text{preprogram}} = 37.4$$

$$M_{\text{postprogram}} = 31.8$$

If the pooled standard deviation was 6.4, the effect size would be

$$g = \frac{(37.4 - 31.8)}{6.4} = .88$$

Effect sizes are best calculated so that a desired effect of the program produces a positive effect size. If an increase in the value of an outcome variable reflects a desired result, then the numerator of the effect size formula would be reversed.

Two questions should come to mind. First, where did $_{\text{pooled}}$ come from? The assumption is made that the true standard deviations of the pretest and the posttest are equal except for random variation. You may recall that the assumption of equal standard deviations is also made when using *t*-tests and analyses of variance. So, the best information we would have in this case would be to use a value halfway between the standard deviation of the pretest and that of the posttest. Suppose that the standard deviation of the pretest was 7.1 and the standard deviation of the posttest was 5.7; the mean of these two numbers is 6.4.*

The second question should be "Is 0.88 good or bad?" To decide, an evaluator needs to know the effect psychotherapy has for treating test anxiety. If an evaluator finds some reports on the effectiveness of treatments for test anxiety, the effect size can be calculated for each report to gain some perspective on the findings of the program being evaluated. In the absence of such information, evaluators can examine summaries of many evaluations. Lipsey and Wilson (1993) summarized more than 300 reviews of evaluations in different areas. They found that the median effect size was 0.44 with the 20th percentile being 0.24 and the 80th percentile being 0.68. In the context of these norms, an effect size of 0.88 would be considered as a large one. As noted above, this form of effect size is known as Cohen's d. Cohen suggested that effect sizes up to 0.3 are small effects, those from 0.4 to 0.6 are medium effects, and those above 0.7 are large effects.

The third question—one that might not have come to mind yet—is this: is it valid if I make my judgments using Lipsey and Wilson's benchmarks? Well, not necessarily. You must retain your common sense. If a workshop was offered to teach how to use a new computer program, a pretest might show just about zero skill in using the program, but a posttest might show a sizeable increase. In this context, we would expect a very large effect size because most participants started at just about zero skill. By contrast, imagine a new medication for people who have had a heart attack that reduces mortality by just 1% among people 45–60 years old. A 1% increase in skill using a program would not indicate a valuable change, but a 1% reduction in death rate due to a heart attack among middle age people is more valuable.

It would be helpful to have a simple way to interpret an effect size; that is not possible. However, finding an effect size does permit one to compare change across similar programs which have used similar, even if not identical, measures of outcome with similar participants. That would have been impossible if effect size formula had not been developed.

Program evaluators who want their work to have an impact on agencies need to be sensitive to the issue of substantively significant change, not just statistically significant change. Determining what constitutes sufficient change is a question for evaluators to address with program stakeholders. Interpretations are more valid when additional information is available: with detailed knowledge of the program, the participants, alternative programs, and the costs of the alternatives, it becomes possible to define meaningful change.

Relating Change to Amount of Service and Participant Characteristics

In educational and training programs, one sign of an effective program is that participants who receive more service change more than those who receive less. The more primary school grades that students complete, the better their reading skills. However, eating more food is related to better health only up to a point; then it leads to problems. The healthiest have found their optimal levels. For the most part, services offered to the poor provide less than the optimal level, although some argue that this is a complex issue (McKnight, 1995). Although correlating service received with outcome may be interpreted in different ways, it can be very useful in certain situations. For example, an important study in the history of medical research followed just this

strategy. In 1835 Pierre Louis (Eisenberg, 1977) reported on his comparisons of the amount of blood drawn from patients with their progress in recovering from pneumonia (then called inflammation). Early nineteenth-century medical theory led to the expectation that the more blood drawn, the better the treatment and the more likely a recovery. Louis measured the volume of blood drawn from patients then followed their condition. He found that their condition after treatment was not related to the volume of blood removed. This finding was an important influence on the medical practice of the time and contributed to the eventual discrediting of bloodletting as a form of medical treatment. Louis was unable to conclude anything about the cause or proper treatment of pneumonia, but he did identify a useless treatment.

Another reason to examine programs even when one is unable to identify the cause of a change is to search for characteristics of the participants that might be related to achieving good outcomes. Do men experience better outcomes than women? Do members of minority groups complete a program as frequently as majority-group clients? These questions can be explored, tentatively to be sure, using a simple research design. When policy-relevant relationships are found between outcomes and characteristics of participants, those variables should be measured systematically in any future evaluations. The finding may even have an immediate impact if a program appears to have a good effect on some segments of the target population but little effect on other segments.

Although it is quite easy to discuss the value of relating personal and service characteristics to success in a program, the statistical method of correlating improvement with other variables is often misunderstood. The most intuitive approach is to subtract pretest scores from posttest scores, call the differences "improvement," and correlate the improvement with any other variables of interest. This seemingly obvious method is not endorsed by methodologists. Raw change scores—the differences between pretests and posttests—should not be used because they are based on questionable statistical assumptions (Meier, 2008). Furthermore, the complex methods suggested to deal with change scores are hard to explain to clients and other stakeholders and are controversial, even among experts. Yet because the lengths of some services, such as psychotherapy, are not standardized and because people receive differing amounts of service, it is of interest to relate outcome to the amount of service received. Fortunately, when evaluators are interested in relating change to characteristics of the program participants or amount of service, there is a valid technique available.

One good way to work with outcome and participant characteristics is to use advanced structural equation modeling (see Raykov, 1999), although most program evaluations do not have the number of cases or other resources to support this analysis. However, there are some analyses of change that are generally possible and that will inform stakeholders. A valid way to relate change to the amount of service received or to some characteristic of the program participants involves *partial correlation*, not raw change scores (Judd, Kenny, & McClelland, 2008). Partial correlation may sound complicated but it is not. First, think about the problem. Those participants who finish the program at good levels are probably those who started at better levels relative to the other participants. For example, students who do well in course X are likely to have higher GPAs prior to the first day of class. The relationship is not perfect, but we think most students expect to find the principle to be true in general. Second, it would not be surprising to find that those with the higher pretest scores turn out to have participated in a program more regularly than those with lower pretests. In college, it is generally true that students with higher GPAs attend class more regularly than those with lower GPAs. In other words, if these principles are correct, we usually expect to find positive correlations among pretest scores, posttest scores, and participation levels. If we want to learn whether degree of participation predicts better outcomes, we need to control statistically for the differences in the pretest scores, that is, differences among participants before they even began, otherwise the correlation between participation and outcome would be large just because both variables correlate with the pretest score. We can control for pretest differences by using a partial correlation between the posttest scores and participation, holding pretest scores constant.

See Example 9.1 in the eResources for hypothetical data for 26 participants in a 20-session job-training program. Imagine that an evaluator wanted to learn how closely attendance was related to improvement in skill level of the trainees. Partial correlation, in effect, statistically removes the effect of pretest differences among program participants on the correlation of the posttest and the level of participation. Virtually no evaluator today would calculate correlation or partial correlation coefficients using hand calculators; instead, we use computers. To do a partial correlation, the variables to be correlated (in this case the posttest score with the level of participation) are identified and the variable that is to be controlled (in this case the pretest

score) is specified. For the data in Example 9.1, the partial correlation is 0.40, indicating a fairly strong relationship between improvement in skill level and number of sessions attended. Training sessions seem valuable in skill level development; the better performance of those who came most frequently is not due to the fact that those who came more frequently started at a higher level.

The partial correlation of 0.40 would be seen as "fairly strong" on the basis of the values of correlations that are found in the social science research literature. As noted above, Cohen (1987) found that 0.10 was small, 0.30 moderate, and 0.50 large. As mentioned above, these values are guides that need to be adjusted for the context in which an evaluator is working.

It is crucial to note that this internal analysis is appropriate only when there are fairly sizable differences among trainees in class attendance. If all trainees had come to 19 or 20 sessions, thus reducing the differences among attendance rates, the partial correlation is likely to be low, as correlations usually are when the range of one variable is small (Zechmeister & Posavac, 2003). In that case, it might appear that attendance did not matter that much. However, the proper interpretation would be that it does not matter much whether trainees attend either 19 or 20 sessions; the low partial correlation would not have implied that 12 sessions would be just as helpful as 20.

Although a positive partial correlation between outcome and level of participation in a program supports the value of the program, it does not eliminate all nonprogram alternative interpretations that might explain why participants improved. The reasons for caution in concluding that a program caused improvements are described in the balance of this chapter.

THREATS TO INTERNAL VALIDITY

Regardless of whether an evaluation shows that participants improved or did not improve, it is important to consider several plausible explanations for the findings. Evaluators want to develop valid conclusions; if some nonprogram influence could be responsible for the findings, we want to know that. The phrase "internal validity" refers to the degree of certainty an evaluator may have concluding that a program caused participants to improve. Many evaluations use these single-group designs. Knowing why these designs fail internal validity criteria guards evaluators from claiming too much.

Actual Changes in the Participants Not Related to Program Participation

Two threats to internal validity refer to real changes that occur in participants due to influences that are not part of the program.

Maturation

Maturation refers to natural changes in people due solely to the passage of time. Children can perform more complex tasks as they get older; people get increasingly tired as they go without sleep; and there are predictable patterns to the development of adults. If an evaluation uses outcome variables that can be expected to change merely with the passage of time between a pretest and a posttest, maturation could be a plausible explanation. In other words, the evaluator may well have found that real changes have occurred during the course of the program; however, the reason for the changes could be that the program lasted six months and thus the participants are six months older and more experienced—not that the participants gained anything from the program. Sometimes people get better over time (e.g., most people with back pain experience improvement with time with or without treatment) and sometimes people get worse over time (e.g., as children get older a greater proportion experiment with smoking and drugs). When using a single-group design, an improvement such as seen with back pain can make a program look valuable over time, but a behavior like drug experimentation can make a prevention program look bad as time goes by.

Finding that maturation has occurred does not mean that it is the only explanation for the changes that were observed. Nor do plausible, alternative hypotheses mean that the program had no effect. Interpretations would be easier if the evaluator could learn how much of the change was due to maturation and how much was due to the program. Using the one-group pretest–posttest design, an evaluator cannot make this distinction. Methods to estimate the change due to maturation involve testing other groups of participants or

potential participants as well as testing at a greater number of time periods, as explained in the next chapter.

History

History refers to events occurring between the pretest and the posttest that affect the participants. For example, an economic recession may make even the most well-designed program to help people find jobs look like a dud. On the other hand, an economic recovery would make a poorly run program look like a winner. These concurrent national economic changes are plausible alternative interpretations of any changes found among the participants of job-training programs.

Some of the same approaches used to account for maturational effects can help isolate historical effects: test additional groups and test at additional times as described in the following chapter. However, history is less predictable than maturation. Unexpected national events that affect the outcome criteria can occur at any time or may not occur at all during a given evaluation. In addition, Cook and Campbell (1979) pointed out that history need not refer only to events affecting all participants; when something unusual happens in a particular program group, this effect is called *local history*. Events such as an argument among staff members, a particularly unusual individual in a therapy group, or a local community disaster cannot be accounted for by any evaluation design except for designs that are replicated in a number of different locations.

Apparent Changes Dependent on Who Was Observed

Three threats to internal validity must be considered when participants are not a random or representative sample of the people who might benefit.

Selection

Participation in many human service programs is voluntary. Self-selected people are different from the typical members of the target population. In the posttest-only form of evaluation, the process of self-selection may mean that the participants were relatively well off when they began the program or that they were especially motivated. The fact that the posttest detects a desirable state reveals nothing about the effectiveness of the program. College teachers most likely to join faculty development programs are often already good teachers. After a faculty development program, most of these teachers will continue to be good teachers. Their competence tells us nothing about the quality of the program—these teachers were better than the typical teacher from the beginning. Observations of their achievements before the program permit evaluators to estimate the misleading effects of self-selection.

Attrition

People differ in whether they will begin a program and they differ in whether they will complete a program they began. The posttest-only design is inadequate when too many participants leave before completing the program. Students drop courses they do not like, clients quit therapy when they learn that personal growth is hard, and medical patients sometimes die. The longer it takes to carry out an evaluation, the more attrition is likely to occur.

The level of achievement observed at the end of a program may indicate how well the program functioned, how good the people were who chose to participate, or how motivated the people were who stayed until the end. As a general rule, those who stay seem to find the program to be more helpful or interesting than are those who drop out. Failing students are more likely to drop a course than are those earning B's and A's. Patients who die were probably among the least likely to have benefitted from a treatment program. Without a pretest, evaluators cannot accurately gauge the effectiveness of a program. A director of a small drug-abuse program remarked informally that his program had a 90% success rate. Upon inquiry, however, it was learned that only approximately 10% of those beginning remain for the full program. Having success with 9 out of the 100 who begin a program is markedly different from having a 90% success rate.

As with selection, the pretest–posttest design handles participant attrition fairly well. By pretesting,

evaluators know who dropped out and how they differed from those who remained. The pretest–posttest design enables evaluators to know when preprogram achievement and participant dropout are not plausible explanations for the level of outcome observed at the end of the program.

Regression

The threat of regression toward the mean is one of the hardest to understand. However, as with basic statistics, many people already have an intuitive understanding of regression toward the mean even though they might be initially puzzled by the idea.

If you think carefully about the following example, you will understand what regression means and why its effects usually go undetected. It is said that after a particularly good landing, pilot trainers do not compliment a trainee, because when such a compliment is given, the next landing is usually done less well. On the other hand, the trainers severely reprimand a trainee after a poor landing in order to elicit a better landing on the next try. Empirically, it is true that complimented exceptionally good landings are often followed by less good ones, and reprimanded bad landings are often followed by better ones. However, let us see why the compliments and the reprimands may have nothing to do with the quality of the next landing.

Imagine learning a complex task. Initially, quality of performance fluctuates—sometimes better, sometimes worse. What goes into the better performances? At least two things: the level of skill achieved and chance. Pilot trainees know that they should touch the plane down with the nose up, but they do not at first sense the precise moment to lower the wing flaps and, at the same time, adjust the elevators to obtain the proper landing angle (Caidin, 1960). Sometimes they make adjustments a little too soon and sometimes a little too late. Because these errors are due partially to chance, the likelihood of a trainee doing everything at precisely the correct instant two times in a row is low. Therefore, for trainees the probability is low for two consecutive good landings—or two consecutive bad ones. In other words, a trainer should not expect two consecutive good landings by a novice regardless of what is said or not said.

The effects of regression toward the mean can be further illustrated with a silly but instructive example. Choose 20 pennies and flip each one six times, recording the number of heads for each penny. Select the pennies that produced the most "excessive" heads—five or six heads can be considered excessive. Reprimand those pennies for producing too many heads. Then flip just those pennies six times again and see if the reprimand worked. If the penny yields fewer heads than it did during the first set, the penny is now behaving in the way it was urged to behave. On the average, the reprimand will appear to have been effective 98% of the time if a penny originally produced six out of six heads, and 89% of the time if the first result had been five out of six heads. The binomial distribution permits the calculation of these percentages (McClave, Benson, & Sincich, 2008).

In summary, regression to the mean warns that whenever the value of a variable is extreme, the next measure of that variable is likely to be less extreme. This principle applies to landing airplanes and flipping pennies as well as to emotional adjustment. If people who currently are the most depressed are followed for three months, it is likely that they will be less depressed as a group. This does not mean that they will be at a healthy level of adjustment. Most likely, they will still be more depressed than the general public. However, some of the transient events that contributed to the depression will have passed, and the most depressed, as a group, will be less depressed than before—with or without therapy.

Is regression a threat to the internal validity of the pretest–posttest evaluation design? Not necessarily. If all the members of a population are tested before and after a program, then regression is not a problem. For example, if all children in a school are given a special reading curriculum, regression will not be a threat to internal validity. However, if only the children reading at the lowest levels on the pretest use the special curriculum, regression will be a threat to the interpretation of the pretest–posttest change. Children score low on a test not only due to poor reading skills, but also due to such random things as breaking a pencil point, having a cold, misunderstanding the instructions, worrying about an ill sister, or planning recess games. A day, a week, or a semester later during a second test, the distracting events will not be experienced by exactly the same children. Generally, these retest scores will be higher than those from the first test for the children who previously scored the worst. This does not mean that all poor-scoring children will improve; it does mean that their average will go up. By contrast, the children who scored the best will on the average do less well.

Regression is often a plausible alternative hypothesis for pretest–posttest change when service programs are aimed at those people who are especially in need of help. Remedial programs are not prepared for everyone, but for those who have fallen behind: reading poorly, earning low wages, or feeling emotional distress. Sometimes a screening test to select people for a program is also used as the pretest in an evaluation plan. For example, students scoring the most poorly on a reading test are placed in a reading improvement program, and a second administration of the same test or a parallel form is compared with the pretest scores. This is a poor practice; regression is a very plausible explanation for at least part of the apparent improvement. On the other hand, if poor readers or troubled clients get worse when regression effects should have led to improvement, then the evaluation can be interpreted. The correct interpretation is that the program is ineffective.

Changes Related to Methods of Obtaining Observations

Two additional threats to internal validity, testing and instrumentation, are generated by the evaluators themselves and by their observation methods.

Testing

The effect of testing refers to changes in behavior due to being tested or observed. The results from repeated uses of the same observation technique may differ simply because respondents have become more familiar with the tool. Ability test scores increase reliably on the second administration for people initially unfamiliar with the test (Murphy & Davidshofer, 2005). People interviewing for jobs gain from the experience and can present themselves better on subsequent interviews.

A second aspect of testing effects is called "reactivity." People behave differently when they know they are being observed. This concept was discussed in Chapter 4; however, it is worth recalling. Clients, patients, prisoners, and schoolchildren will be affected when they know someone is observing their behaviors or asking them about their opinions or feelings. Observation techniques vary in how reactive they are.

The pretest–posttest design is clearly weak in the control of testing effects. If participants were unfamiliar with observation procedures, scores might change on the second test. The kind of change that should be expected due to repeated observation or testing does not seem clear except for ability and achievement tests; on these improvement is expected.

Instrumentation

Instrumentation refers to the use of measurement procedures themselves. Most college instructors know that they are not totally consistent when grading essay examinations. Standards can change as the instructor becomes familiar with how the students answered a question. The standard may become higher or lower. If measures in an evaluation are not highly objective, it would be wise not to score the pretest until after the posttest is administered. Then the tests can be shuffled together and scored by someone who does not know which are pretests and which are posttests.

If the measures require observations that must be made before and after the program, the examiners may become more skilled as they gain experience. Thus, the posttests may go much more smoothly than the pretests. If so, a change in instrumentation becomes a viable alternative to any conclusion the evaluation might imply. In such situations, examiners who are highly experienced before the pretest is administered are the most effective.

Effects of Interactions of These Threats

In addition to these seven threats to internal validity, interpretations can be confused by the joint influence of two threats. This could occur if parents seek special educational opportunities for their children (an example of self-selection) because their children are developing more rapidly compared with the children of parents who do not seek special educational opportunities for their children (an example of different rates of maturation). This effect is called a *selection-by-maturation interaction*. It is generally especially hard to predict

the direction and size of effect of interactions, however.

Internal Validity Threats Are Double-Edged Swords

When examining a comparison between means, it is important to consider whether an apparently positive effect could have been caused by a threat to internal validity increasing the difference between the pretest and the posttest. Although attention is usually focused on these threats masquerading as program effects, it is also quite possible for these threats to hide a program effect. Therefore, the lack of a program effect, when the samples are large and the measures reliable, could be due to an ineffective program or to an uncontrolled threat to internal validity serving to reduce the difference between pretest and posttest. Examples include an economic downturn, participants being tired at the posttest, the most successful participants being unavailable for the posttest, initially good-scoring participants regressing toward the mean, and more sensitive observation procedures detecting dysfunction that was missed by a less sensitive pretest.

In the next chapter, several design enhancements will be presented as ways to eliminate plausible explanations of change that are not related to the program. Only when such explanations are implausible can good outcomes be attributed unambiguously to the program. Likewise, good experimental control would provide better understanding of apparent program failure.

CONSTRUCT VALIDITY IN PRETEST–POSTTEST DESIGNS

Construct validity refers to the concept (or construct) that a method of observation actually measures. When evaluators use participant self-report surveys in pretest–posttest designs, it is necessary to examine whether the program might lead to changes in how participants label their own problems and strengths. Veterans suffering from posttraumatic stress reported that a program was very helpful, but described themselves as functioning less well and experiencing more symptoms than they did before completing the program (Spiro, Shalev, Solomon, & Kotler, 1989). Was the program harmful? It appeared that the program led them to a more accurate recognition of their problems and a greater willingness to acknowledge them. The program was helpful, not harmful. Evaluators have distinguished different kinds of change; to differentiate among them they are called *alpha*, *beta*, and *gamma* changes (Arvey & Cole, 1989; Millsap & Hartog, 1988).

An *alpha* change is a real change in the behavior of interest. Identifying the cause of the change is a question of internal validity, as discussed above. *Beta* changes occur when respondents change their understanding of the meaning of the variable. For example, exposure to a really good golfer may lead a person to rate her own skill at golf less favorably than when playing only with friends. People with better insight into what level of skill is possible may see themselves as functioning less well than they did when they completed a pretest; this change appears numerically as a loss, but the change may indicate a more accurate self-appraisal. *Gamma* change refers to differences between the pretest and the posttest that are due to a reconceptualization of the meaning of the variable being measured. For example, a high school teacher might not see his physical symptoms as resulting from stress; once he does, his ratings of his own stress level may change because his understanding of the meaning of his symptoms has changed. Data taken from records may also be subject to these artifacts when official definitions change over time. Variables that can be measured objectively would not be subject to *beta* and *gamma* changes. When a variable could mean different things to different participants or different stakeholders, we say that the variable lacks construct validity.

For evaluations requiring a pretest that includes self-reported behaviors, attitudes, or intentions, several strategies can be used to minimize these threats to construct validity, including (1) distinguishing information gathered to evaluate programs from that needed to make treatment decisions, (2) implying that what respondents report will be validated in some way, (3) using interviewers who are experienced with both the program being evaluated and the problems that the respondents have, (4) providing explicit reference groups for respondents to compare themselves with (e.g., "How do your leadership skills compare with other managers in your division?" rather than "Rate your leadership skills from 1 [poor] to 10 [excellent]"), and (5) using behavioral anchors rather than evaluative terms (e.g., "I am so depressed that I cannot accomplish anything" rather than "Very depressed"). In some cases, retrospective pretests may be better than tests given before the program begins because after experiencing the program, participants have a better understanding of themselves (Robinson & Doueck, 1994). However, there is evidence that sometimes participants inflate their

own program success by reducing their preprogram rating of themselves (Taylor, Russ-Eft, & Taylor, 2009).

OVERINTERPRETING THE RESULTS OF ONE-GROUP DESIGNS

When evaluators use a simple design, there may be a temptation to seek to compensate for the design's weaknesses by measuring many variables. When this happens, the chances of a Type I statistical error increase. People not trained in research methods seldom appreciate the high probability that patterns of relationships will be observed among some variables just through sampling error. Many people grossly underestimate the probability of random coincidences, but instead think that causal connections have been observed (Paulos, 1988). Type I errors refer to statistically significant differences between pretests and posttests even when no change occurred. Type I errors occur for exactly the same reason that flipping 20 coins six times each will appear to identify some coins with a tendency to produce a surprisingly high proportion of heads. We can recognize that random variation was responsible simply by repeating the set of six flips. Unfortunately, when performing evaluations new observations cannot be obtained as easily as flipping coins several additional times.

A second reason why Type I errors are not discovered is related to hindsight bias. Hindsight bias is the tendency to believe we could have anticipated something *after* we are aware of what in fact happened (Bryant & Guilbault, 2002). In program evaluation, hindsight bias refers to the tendency to make sense of almost any set of data after having examined it. Once a puzzled graduate student (actually, the second author of this text) presented a professor with findings that seemed markedly at variance with the professor's research. The professor examined the columns of correlations and quickly interpreted the findings in a way that coincided nicely with the previous studies. In order to do that he developed novel, elaborate interpretations that had not been part of the theory. A week later the student discovered that the computer program had malfunctioned and the initial results had been mislabeled. An alternate program produced results that were quite compatible with the professor's work. When faced with a page full of means for different measures or with a page of correlations, creative people can make sense of them—even when the variables are mislabeled and the results are in the wrong direction!

These problems, ignoring the possibility of Type I errors and falling prey to hindsight bias, are most likely to affect interpretations when many variables have been measured and when evaluation designs have few ways to control for threats to validity. Evaluators often have access to many demographic variables and can extract many process and outcome variables from agency records. Without a clear impact model indicating how the program processes are expected to work, it is easy to succumb to the temptation to correlate everything with everything else. Then, evaluators or their clients may read interpretations onto the results. Novel findings are best treated as very tentative, especially if they appear in the context of a large number of analyses that were not suggested by the theory underlying the program. All evaluation designs can be overinterpreted, not just single-group designs; however, the tendency seems more likely because this design is descriptive and does not support the inference that participation in the program causes desired outcomes.

USEFULNESS OF ONE-GROUP DESIGNS AS INITIAL APPROACHES TO PROGRAM EVALUATION

When there are previously specified desirable levels for the outcome variables (benchmarks), and when the participants do not drop out during the program, the pretest–posttest design may be sufficient to document the program's success. However, even when benchmarks are not available, the reader should not gain the impression that these approaches do not have any legitimate uses. These single-group designs are less intrusive and less expensive and require far less effort to complete than more ambitious methods. Thus, these designs can serve important functions in an evaluator's toolbox as first steps in planning a more rigorous program evaluation. When standards are not available, the purposes of single-group evaluations include (1) assessing the likely usefulness of a more rigorous evaluation, (2) searching for promising variables correlated with success in the program, and (3) preparing an agency for more controlled evaluations in the future.

Assessing the Usefulness of Further Evaluations

Before embarking on a rigorous evaluation that controls for many plausible alternative interpretations, a single-group design would serve to show whether there is any improvement to explain. If participants fail to get jobs for which they were trained, it is likely that no further study is needed. The planned outcomes were not achieved; no statistical analysis is needed. If the rehabilitation program participants finished at desirable levels of function, a pretest–posttest design might show that they improved while in the program. This finding may well satisfy some of the stakeholders. Moreover, finding improvements might justify a more complex evaluation.

Correlating Improvement with Other Variables

The use of the pretest–posttest design and partial correlations permits relating improvement to the amount of service and to characteristics of the participants. The mean level of improvement could be acceptable, but if outcome measures show large standard deviations, it might be that some subgroups of participants benefit, while others benefit little. That question should be explored. On the other hand, a favorable outcome at the end of a program may really indicate a general improvement for many participants. Perhaps the improvement was due to the impact of the program; further work would be merited.

Preparing the Facility for Further Evaluation

A third reason for conducting an evaluation using only the participants is to help the staff understand the idea of evaluation. As described in earlier chapters, service providers—from paraprofessional counselors to teachers to physicians—are beginning to recognize that their work can and will be evaluated, even though the quality of service is hard to quantify. If evaluators begin their work with less threatening approaches, they have better chances of leading service providers to see the usefulness of evaluation and to value the contributions of evaluators.

The methods described in this chapter are among the least threatening. There is no possibility of learning that another program achieves the same level of success at a lower cost. When staff and managers fear that another program is more efficient, the resulting anxiety can lead to hostility and limited cooperation with an evaluator. Just this type of reaction occurred during an internal evaluation of a hospital rehabilitation program. The program director initiated the evaluation and showed remarkable openness to evaluation, making comments such as "Maybe we don't help people as much as we think we do—we should know that" and "If a patient is institutionalized after leaving here, we have failed." However, when the evaluators proposed increasing the internal validity of the evaluation by gathering observations from a neighboring hospital without a similar rehabilitation program, the director seemed to change his position. Such a comparison could conceivably have shown that similarly diagnosed patients without a program did as well as those in the expensive program. Then he said that the additional data were unnecessary since the value of this type of program was well known and fully documented. The expanded study went on because no controlled studies could be found supporting his assertion, a situation that is surprisingly familiar to those who carefully search for support for some widely accepted practices in medicine (see, for example, Eddy, 1990). The director mellowed when it was learned that the neighboring hospital had very few patients with the appropriate diagnoses and thus no information could be forthcoming that would question the need for program he directed. Cooperation and cordial relations returned.

Rational use of resources requires that ultimately such comparative studies be done. In medicine, psychology, and social work much attention is being paid to "evidence-based" practices; however, many observers report that the evidence cited is too often equivocal (Raines, 2008; Weiss, Murphy-Graham, Petrosini, & Gandhi, 2008). Careful evaluations should not be done in a way that causes a staff to feel threatened. Internal evaluators who try to insist on methodologically rigorous evaluations may find that they lose their influence with the staff. Beginning with nonthreatening methods permits evaluators to gain the confidence of staff members and managers. This strategy is not meant to imply that rigor is unnecessary; however, using less rigorous evaluations at first can be useful in earning the trust of the staff.

CASE STUDY 9

Although spiritual needs would not be the first thing to come to mind for many people when considering how to serve homeless people with substance abuse problems, for about 20 years, the Ignatian Spirituality Project (ISP) has been conducting spiritual retreats for homeless adults in a number of U.S. cities. As one might imagine, this kind of program regularly has more needs than financial resources, so hiring a program evaluation firm was not a high priority. But Dn. Joe Ferrari, PhD, a St. Vincent de Paul Professor of Psychology at DePaul University in Chicago, was the right man in the right place at the right time to connect appropriate people and resources (see Evaluator Profile 9 for more specifics). As is intended with all the case studies, this one illustrates important points from this chapter—in this example, particularly noting elements related to the evaluation of a single group as opposed to multiple groups that included a control condition. In addition, however, a major theme through Dr. Joe's work with ISP illustrates a pervasive point about working with most stakeholders—evaluations require trust, and trust takes time.

Meeting Needs

The broader base of needs is very clear—it is estimated that there are up to 400,000 homeless adults on any given night in the United States, many of whom have disabilities of some form and therefore would benefit from appropriate services. Although many may be in temporary circumstances, a substantial number are chronically homeless. The reality that spiritual needs are seldom considered as an appropriate focus for homeless men and women with substance abuse problems is a clear example of common societal attitudes and a related part of homeless people's experience—many people would prefer not to have to be aware of their existence, let alone their needs, and homeless folks are generally well aware of this fact. Some research supports an awareness among homeless people of their spiritual needs, but this point is not generally well known. Although not noted explicitly in the evaluation report, one implicit point about ISP's provision of spiritual retreats for homeless people is that taking the spiritual needs of homeless people seriously communicates something very important about the value of those people both to the homeless themselves and to society more generally. Further, as will be seen in the following, addressing spiritual needs may also have effects on other needs that are not necessarily spiritual.

Implementation

As the retreats are weekend events (Saturday morning to Sunday afternoon), the report describes what was observed by the evaluation team, whereas monitoring implementation for ongoing programs typically requires much more extensive processes. For example, the first event at 9:00 am on Saturday is an ice breaker followed by a prayer. Various individual and group activities follow with an obvious theme of considering God's relevance to one's life and taking time for silence and reflection. Meals and overnight accommodations are provided for these same-sex events. Toward the end of Sunday's activities, a group discussion provides an opportunity for participants to summarize intentions or to reflect on their experience.

Stakeholders

Most evaluations include some stakeholders who are reluctant, even resistant, to some aspect of the evaluation. Dn. Joe is clear that the evaluation of ISP retreats took a long time to develop. The ISP director initially contacted Dn. Joe, based on a previous association, regarding some analysis of data that ISP had already collected. One step at a time, Dn. Joe and some of his students became more involved with the project—first with formal data analysis, then summarizing comments from a reflection session, then publishing the results of those summaries, and on to more structured evaluations, a result of growing trust over time. Although not stated explicitly, it would not be surprising that ISP staff were motivated to protect their retreatants from potential harm, given common attitudes and behavior toward the homeless. But clearly the staff and the participants, the other major category of stakeholders, grew to trust the evaluation team.

Side Effects

Especially in this case, the distinction between side effects and outcomes depends on one's perspective. Although one could describe having a spiritual retreat result in decreased loneliness and increased hope as positive side effects, the evaluation team, focusing on ISP's obvious intention for positive results, understood more broadly than narrowly, listed them as outcomes. No negative side effects were noted.

Improvement Focus

As with several aspects of this evaluation, the focus on improvement is somewhat subtle in the report—blending summary information about the participants and their experiences with gentle suggestions for ways to increase the scope of the evaluation and, therefore, the basis for feedback to ISP and eventual improvement of the program.

Outcomes

As noted above, the outcomes of decreased loneliness and increased hope over time, which were greater among women and African-Americans than in other groups, were a positive result documented by the evaluation team. Although there were no structural research design elements to support attributing the effects to the intervention of a spiritual retreat as opposed to alternative explanations like the participants being invited to a novel experience that was likely in a more pleasant setting than some shelters, the documentation does provide an important baseline for future evaluations as well as the clear suggestion that the retreat itself led to good outcomes. One important point about all evaluation results is that, although they are often not completely convincing evidence that the program must have caused the benefits, in the absence of clear evidence to the contrary they provide a reasonable hope that the program is resulting in good effects. Often that is enough to convince stakeholders to continue to invest time, money, and effort into the program.

Nuances

The evaluation reports touched on spiritual needs as well as the way that when people are involved in religious organizations, they often experience important elements of social support, which is helpful to people who are lonely. Although the evaluation did not explicitly test this assumption that participation in the retreats led to reduced loneliness because of social support, a general sense of human connection was clearly implied, and such a possibility is a reasonable hypothesis for more rigorous scrutiny in the future.

Summary

As the text noted, evaluations with one group have important limitations, and the conclusions that can be drawn from such evaluations generally have lower confidence and precision than other designs. Acknowledging these limitations, a corresponding point to emphasize is that many programs, for a variety of reasons, are in a position where a single-group evaluation can provide some insight and the only alternative is to have no evaluation. It is clearly important to understand the inherent limits of these designs, but it is equally important not to ignore evaluations with limits simply because they are not perfect.

Sources: J. Ferrari (personal communication, October 26, 2017); Ferrari, J. R., Drexler, T., & Skatt, J., *Psychology,* 6, 210–216, 2015; Ferrari, J. R. et al., *J. Addict. Res.,* 1, 1–4, 2017.

EVALUATOR PROFILE 9

Dn. Joseph Ferrari, PhD—Called and Formed

Rev. Dr. Joe Ferrari, a Permanent Deacon in the Catholic faith, notes that he was "called and formed" to his current roles, which cover a very wide range. In addition to serving the people of the Diocese of Joliet, Illinois, he was the founding director of an MA/PhD program in Community Psychology at DePaul University in Chicago, IL. He is an extremely productive, multiple award-winning scholar, a St. Vincent de Paul Professor, and the Editor of the *Journal of Prevention and Intervention in the Community*. Still, he notes that his path to the present was anything but direct.

As he worked his way step by step toward a doctorate, he was particularly drawn to community psychology with his interests in applied matters that made a difference for individual people and communities like increasing blood donations and seat belt use. Yet he entered a doctoral program in social psychology because the Department Chair said, "Blood donations? That sounds like social psychology. Besides, we don't have community psychology here." After several legs of the journey in New York, he set off for Chicago for a big adjustment to a much larger program than when he himself comprised half the psychology department. At DePaul University, writing grants, publishing, teaching, and editing still did not keep him from his passion of offering students a chance to help people really make systemic changes in life. So an observer might think it is hardly surprising that one of his program evaluations examined the Ignatian Spirituality Project's spiritual retreats for homeless—a particularly striking example of helping others.

Dn. Joe, as he likes to be known, is clear that the destination is not always known at the start, noting that "We never know where life will take us—the directions we are called, the events that form us." With reference to the Source of how things develop in our lives, he explains, "PhD goes at the end of my name, because I earned that. But Dn. goes before my name, because that is what I am called to be."

Source: J. Ferrari (personal communication, October 24, 2017).

SUMMARY AND PREVIEW

This chapter emphasized two major points. First, a sympathetic understanding of the needs of the users of an evaluation may indicate that a simple research design serves quite well to meet the information needs at a particular time. Second, single-group designs cannot answer many questions that are critical to stakeholders because many interpretations of findings remain plausible. These plausible interpretations are called threats to internal validity.

The next chapter contains several approaches to clarify the interpretation of outcome evaluations. These procedures require making additional observations of (1) the program group, (2) nonprogram groups, or (3) variables not expected to be affected by the program.

STUDY QUESTIONS

1 Not many years ago, when a child had frequent bouts of sore throat and earache, family doctors recommended removing the child's tonsils. Typically, the child's health improved afterwards. Parents and physicians attributed the improved health to the operation. What threats to internal validity were ignored?

2 Some 50 elementary school children witnessed a brutal murder on their way to school in a suburb of a large city. The school officials worked with a social worker, a school psychologist, and others in an attempt to help the children deal with their emotions, to avoid long-term ill effects, and to help parents with their children. A year after the tragedy, there appeared to be no serious aftereffects. What assumptions are necessary to claim that those efforts were useful?

3 Politicians are notorious for attributing any improvements in governmental and economic affairs to their own efforts and any deterioration to the policies of others. What threats to internal validity are most likely to make such interpretations invalid?

4 Why can a smoking cessation clinic be evaluated using a one-group design?

5 Prepare some examples of pretest–posttest evaluations that would appear favorable because of threats to internal validity. Also, prepare some hypothetical examples of evaluations that would appear ineffective because of threats to internal validity.

---

* Technically, we pool variances, s^2, and then calculate the standard deviation from the pooled variance. When finding pooled standard deviations for a pretest–posttest design, the use of the average of the two standard deviations is sufficiently precise for most evaluations.

ADDITIONAL RESOURCE

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Boston, MA: Houghton Mifflin.

Although an older text, this is the standard research design resource for program evaluators. It is the successor of the classic work by Campbell and Stanley (1963) and the more recent standard by Cook and Campbell (1979). The first two authors are highly respected methodologists whose writings are relevant to program evaluation. Shadish has been the president of the American Evaluation Association and has published widely on program evaluation. Cook is, perhaps, the most eminent living social research methodologist. The late Donald Campbell has been associated with more innovations in social research methods than anyone has tried to count. Threats to internal validity are discussed in detail in Chapter 2 and other chapters.

10 Quasi-Experimental Approaches to Outcome Evaluation

Very simple research designs can answer some important questions for stakeholders. However, Chapter 9 stressed that those designs seldom permit evaluators to identify the cause of changes in program participants in an unambiguous fashion. Whenever program administrators or government officials want to know whether programs cause such changes, evaluations of greater complexity must be conducted. In order to show that something causes something else, it is necessary to demonstrate (1) that the cause precedes the supposed effect in time; (2) that the cause covaries with the effect, that is, the more of the cause, the more of the effect; and (3) that no viable explanation of the effect can be found except for the hypothesized cause. It is easy to satisfy the first criterion, and the second is not that difficult to test either. Both can be demonstrated using the methods described in Chapter 9. The third, however, is much more difficult because the threats to internal validity must be eliminated as viable alternative explanations for any effect found.

In this chapter several research designs are presented that possess greater internal validity than the simple designs already described. The validity of outcome evaluations seeking to test causal hypotheses can be increased by (1) observing the participants at additional times before and after the program takes place or begins; (2) observing groups of people who have not experienced the program; and (3) using a variety of variables, some expected to be affected by the program and others not expected to be affected. Designs using these methods were called quasi-experiments by Campbell and Stanley (1963); although these methods do not provide the high level of clarity provided by true experiments, quasi-experiments control for many biases and can thus yield highly interpretable evaluations—if carefully planned and applied in appropriate settings. One important motivation for using such designs is that they generally require fewer resources, such as time and money, than true experiments, and thus are more practical for many evaluations.

INCREASING THE NUMBER OF TIMES OBSERVATIONS ARE MADE

Making numerous observations at different times can help resolve the problem of distinguishing random changes from time period to time period from systematic changes in the outcome being measured. All outcome variables show variation over time. For example, the number of crimes on a given day or week is not a constant, nor is it a uniformly decreasing or increasing value. Although there are some predictable influences on the number of crimes committed (e.g., cold weather inhibits violent crime, warmer weather seems to encourage it) (see Anderson, 2001), there seem to be no systematic causes of most changes in the day-to-day rate. Likewise it is impossible to predict day-to-day fluctuations in the indexes of the value of investments; nevertheless, commentators in popular media often talk as though such causes were well known. After an increase, one hears comments such as "Investor confidence led to …" After a drop, comments such as "Uncertainty about …" are heard. Such comments are made *after* the changes were observed. There are trends to explain, but more information must be gathered before offering an explanation. Without information on the extent of day-to-day variation, one cannot know whether there is anything that needs to be explained. A procedure called *statistical process control* is used in manufacturing (Mitra, 1998) and increasingly in human services (Carey, 2003; Henderson et al., 2008; Spiring, 1994) to identify when a particular observation is sufficiently different from the general pattern to merit attention.

The effects of changes in laws can be evaluated by examining time series because a law applies to everyone in an area (not a selected sample of people) and a law will not be reversed except in very unusual circumstances. A law was changed permitting states to divert federal funds from the intended use (supporting

foster care for children suffering abuse or inadequate care in their homes) to an alternative use (helping the biological families that had given inadequate care to their children). States received a waiver to permit this. The objectives were twofold: (1) the family problems might be corrected through counseling and education and (2) it would be unnecessary to remove children from their homes. The number of children placed in foster care per month is a variable routinely tracked by state agencies. Permission to divert funds from the original intended purpose to attempts to correct problems thereby averting foster care placements was granted at a specific time.

Therefore, it was possible to track the number of children in foster care and the number entering foster care before the change and then after the change. If permission to divert these funds led to the desired outcomes, both variables should fall. And they did. This number shows considerable variation, sometimes dropping by as much as 25% from month to month and at other times increasing by a similar percentage. One might be tempted to explain a particular drop; however, one should not. But the trend line of placements showed a drop after the waiver of rules was granted and the trend stayed down, although it still showed month-to-month variation (James Bell Associates, 2009). To summarize, the critical point is that the desired outcome, fewer out-of-home placements, dropped after the wavier was granted and it stayed down even though month-to-month variation was still occurring.

A time series was also used to examine the effect of an HIV (AIDS) testing law in Illinois, which required couples to take an HIV test before being granted a marriage license (McKillip, 1991). Examining the number of marriage licenses granted per month for 116 months showed that the law was associated with an abrupt 14% drop in the number of marriage licenses obtained in Illinois. Adjacent states, without similar laws, experienced a corresponding increase in marriage licenses, but other midwestern states not adjacent to Illinois showed no change in the number of licenses. It appeared clear that a sizable portion of Illinois couples were obtaining licenses in adjacent states to avoid going to the trouble of being tested.

In recent years considerable attention has been given to the use of information across many time intervals. Both economists and behavioral therapists as well as evaluators use time-series designs. Economists have used information over many time periods, seeking to learn the effects of policy changes on such variables as income level, industrial output, and employment. On a very different scale, behavioral researchers utilize objective measurements of single individuals over many time periods, during which various therapeutic interventions are begun and stopped. The strategy of obtaining baseline measurements before an intervention and documenting both change and the maintenance of change has been adopted by program evaluators.

The data used by economists and behavior analysts are more similar to each other than might be supposed on the basis of the focus of their work. Whether referring to a country or a single child, the researcher obtains one value or makes one observation of each variable during each time interval studied. For example, for economists there is only one GDP and one inflation rate for each time period. For behavior analysts there is only one index of a specific abnormal behavior per time period. Thus in both cases, writers can refer to single-subject research.

Collecting data over many time periods is a way of meeting some of the internal validity challenges described in the previous chapter. Because maturation effects can be traced during the time periods before and after the intervention, the likelihood of confusing a program's effect with maturation effects is greatly reduced when using a time-series design rather than a pretest–posttest design. Similarly, the effects of history will be more easily detected using a time-series design than is feasible in evaluations using observations at only one or two time periods. By relating changes in dependent variables to historical events, it is possible, at least at a qualitative level, to distinguish the effects of the program from the impact of other influences.

A time-series approach to program evaluation minimally includes (1) defining a single unit and (2) making measurements (3) over a number of time intervals (4) that precede and follow some controlled or natural intervention. Program evaluators, unlike economists, use time-series designs almost exclusively when a planned change has occurred at a specific time. The design is often called an *interrupted time series* (Cook & Campbell, 1979) and the evaluator's job is to learn whether the interruption—that is, the introduction of a program—had an impact.

Patterns of Outcomes over Time Periods

There are a number of possible patterns observable in graphs of a program's outcome plotted over time periods; Figure 10.1 illustrates some of these possibilities. The axes in each part of the figure are the same: time intervals are plotted on the horizontal axis and the magnitude of the outcome variable is plotted on the vertical axis. Panel A illustrates no effect of the intervention; there appears to be no out-of-the-ordinary change in the values of the criterion after the program was introduced. By contrast, Panel B illustrates what is usually the most hoped-for finding in an interrupted time-series analysis. The graph shows a marked increase from a fairly stable level before the intervention, and the criterion remains fairly stable afterwards. This result may be expected from an effective training program that increased the skills of employees or from the introduction of laborsaving devices. Clearly, the design can be used when the intervention is designed to decrease the value of variables (such as accident rates, failure rates, or the number of children taken from their biological parents) as well as to increase them.

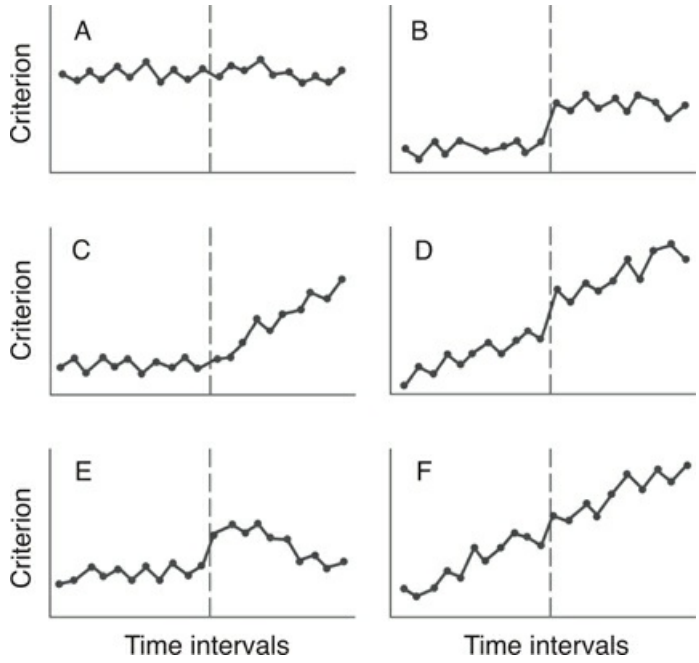


Figure 10.1 Some illustrative possible patterns of a measure of program success plotted over time. Time of the program or intervention is indicated by the dashed vertical line.

Panels C through F all show some increase after the intervention, but their interpretations are much less clear. Panel C shows an increase in slope after the intervention. The variable being measured began to increase over time after the intervention; however, there was no immediate impact as in Panel B. An influence such as literacy training or improved nutrition, whose impact is diffuse and cumulative, might produce a pattern of findings as in Panel C if economic well-being were the outcome variable. In Panel D, it appears there are two patterns superimposed on each other. There is a general trend of increase over time. And there is an additional slight increase due to the intervention.

In Panel E there appears to be an effect due to the intervention; however, it seems temporary. Many new programs are introduced with much publicity, and deeply involved staff members believe in the value of the program and work hard to make it effective. However, once the program is part of the regular procedure and the enthusiasm of the staff has diminished, the outcome variable returns to its former levels. In other cases the novelty of a program might lead to temporary effects.

A possible pitfall for an evaluator could exist when a pattern such as that illustrated in Panel F is found. Here a steady increase over time is observed. A naive approach to the analysis would be to contrast the mean of the time periods before the intervention with the mean of the time periods after. These means will be markedly different and a *t*-test would be statistically significant, but that does not show that the intervention was responsible for a change; it looks like a historical trend, not the effect of the intervention.

Analysis of Time-Series Designs

A discussion of rigorous analyses of interrupted time-series data is beyond the scope of this text, but a recent comparison of a randomized experiment with an interrupted time-series design found very similar results, supporting its value and use (St. Clair, Cook, & Hallberg, 2014). The methods commonly accepted are described by Reichardt and Mark (1998). Some simpler approaches to analyzing a time series may be completed with relative ease.

When the Intervention Can Be Implemented and Then Removed

Some interventions can be started and then temporarily suspended. After establishing a baseline, behavior therapists introduce an intervention that is supposed to reduce the frequency of a problem behavior. Suppose that the intervention appears to be effective: the problem behavior decreases. After several observation periods, the intervention is removed. If the rate of maladaptive behavior increases, it appears that the intervention had an effect—the change was not due to just maturation or history. If the intervention is reintroduced and maladaptive behavior decreases again, it is quite safe to conclude that the intervention is effective. This approach is a standard ABAB design used by behavioral therapists (e.g., Horton, 1987).

When the Intervention Cannot Be Removed, but the Effect Is Large

In other situations, interventions cannot be removed and reinstated. Government policies and organizational changes are not going to be suspended to permit the effect of the change to be evaluated; however, if the effect is large, the impact of the change might still be obvious. The medical education research literature shows that resident physicians—physicians receiving their specialization training in hospitals—order more diagnostic laboratory tests for patients than do licensed, experienced physicians (Rich, Gifford, Luxenberg, & Dowd, 1990). To learn whether this general finding held for a particular hospital, a graph was made of the number of laboratory tests per month beginning four years before a major expansion of residency programs and continuing for two years afterwards (see Figure 10.2). Because neither the hospital's size nor its occupancy rate (percentage of occupied beds) changed systematically during that six-year period, the influence of those historical effects was ignored. One striking observation was the regular increase in the number of laboratory tests during years before the residents began to provide care for patients: there was an average of approximately 500 additional tests each month. However, the increase that occurred between the last month before the program was expanded and the first month of the expanded residency program was more than 10,000 tests. Because the monthly increase resumed after this big jump—that is, the graph resembled Panel D in Figure 10.1—it seemed safe to attribute the increase to the test-ordering practices of residents who were then ordering the bulk of the tests.

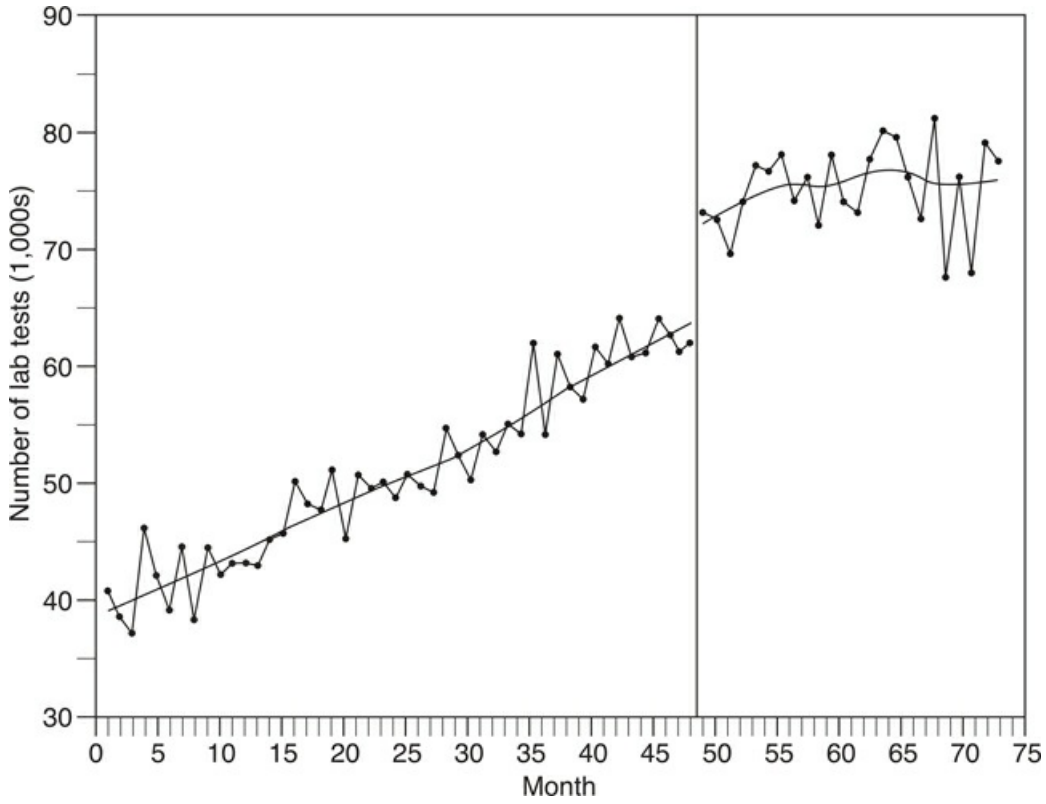


Figure 10.2 Number of laboratory tests before and after medical residents began to provide care to patients of the hospital. A Lowess line was found for the 48 months prior to the beginning of the residency program and for the 24 months after the program began. Note the jump between the 48th and 49th months.

These data could have been analyzed using sophisticated methods. Instead, the plots were simply smoothed for the 48 months prior to the increase in the number of residents and, *separately*, for the 24 months after the increase in residents. Figure 10.2 shows the jump at the point when residents began. The smoothed lines in the figure are called *Lowess lines* and can be made in SPSS after making a scatterplot from the data. It is important to remember that smoothing is done in two steps: first for the 48 months before the residents started, and second for the 24 months after they started. The impact of the residents is hard to ignore in the figure. (Indeed, the program director instituted procedures to reduce the use of tests by residents; the effect of these procedures seems to be reflected in the last 12 months in the figure.) Because the effect was so large, the interpretation seemed clear. If the effect had been smaller, removing and later reintroducing the intervention would have made the interpretation compelling. But since a community hospital does not change into a teaching hospital without great effort, this organizational change was not going to be removed.

A time-series graph can be smoothed without using a computer. See Example 10.1 in the eResources for sample data and graphs.

When smoothing techniques are described the first time, some people object, saying that the data are being changed. Smoothing a graph is no different in principle from finding the mean of a set of numbers. We seldom list our raw data in reports; instead we report means and standard deviations because the mean reveals a typical value better than a list of original data points. In the same way a smoothed graph can reveal a pattern in a graph better than a graph of the raw data.

Although evaluators want as many observation periods as possible, simply getting several is often a great help in interpretation. The introduction of intensive case management for chronic schizophrenic patients was evaluated with three pre-intervention and three post-intervention observations of outcome variables such as rehospitalizations, medical appointments kept, and psychological mood (McClary, Lubin, Evans, Watt, & Lebedun, 1989). These six observation periods made a compelling case that the case management intervention was successful in improving the lives of the patients.

OBSERVING OTHER GROUPS

Nonequivalent Control Group Designs

The simple time-series design increases the interpretability of an evaluation by extending the periods of observations over time. Another approach is to increase the number of groups observed. If the pretest–posttest design could be duplicated with another group that did not receive the program, a potentially strong research design would result. As long as the groups are reasonably comparable, nearly all the internal validity tests are satisfied by this design, called the *nonequivalent control group design*. Nonequivalent control groups may also be called *comparison groups*. The phrase "nonequivalent" is used to indicate that the two or more groups did not begin as single set of potential participants who were then randomly assigned to conditions, as generally happens in formal experiments.

A larger improvement between pretest and posttest is expected for the program group than for the comparison group, as illustrated in Figure 10.3. A statistical tool to analyze such data is the two groups by two time periods analysis of variance, with repeated measurements over time periods (Aspelmeier & Pierce, 2009). Other statisticians call this a *mixed design*, because one element of the analysis is comparing what happens to the same participants over time, called "within subjects," and another element compares what happens to one group in contrast to another group, called "between subjects." The analysis mixes the first element with the second. If the program is successful and the group means follow the pattern in Figure 10.3, the analysis of variance would reveal a significant interaction between group and testing period.

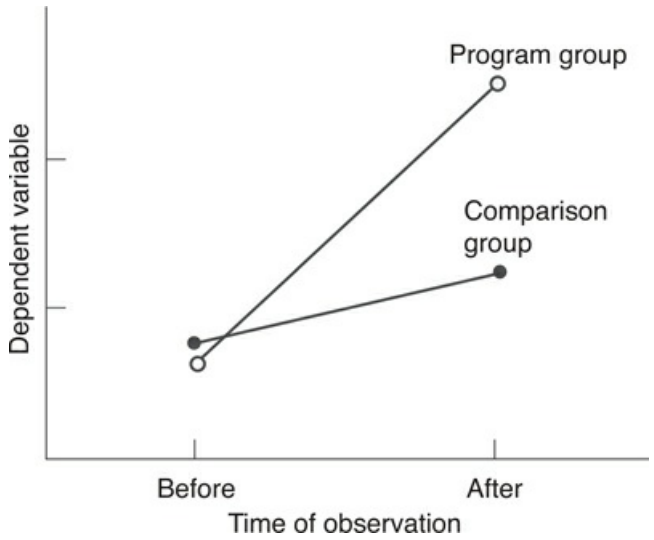


Figure 10.3 Hypothetical results of a nonequivalent control (comparison) group design.

Other statistical analyses could also be used: Reichardt (1979) showed how to use analysis of covariance using pretest scores to control for initial differences among program participants. However, he also showed that depending on the pattern of the group means and how the comparison group is chosen, this approach to analysis can bias the results in different directions. His discussion is too detailed to be summarized; however, the use of analysis of covariance with experiments is discussed in the following chapter. It is very important for evaluators not to be misled by comments referring to analyses of covariance as "correcting for differences between the groups at the pretest"; this comment is often heard, but it is wrong, as will be illustrated just a little later in this chapter.

Before beginning any statistical analysis it is always wise to inspect the data carefully. Ideally, we would want to find that the standard deviations associated with the means in Figure 10.3 are similar. Furthermore, the figure shows that the means before the program were nearly equal; if the pretest means are quite different, analyses of nonequivalent control group designs may be misleading.

Note the positive features of this design. Including the comparison group permits the isolation of the effects of the program from the effects of several plausible interpretations of change. Because the comparison

group would have been tested at the same times as the program group, both groups have had the same amount of time to mature. Historical forces have presumably affected the groups equally. Because both groups would have been tested twice, testing effects should be equivalent. Finally, the rates of participant loss between pretest and posttest can be examined to be sure that they are similar. Nonequivalent control groups are especially useful when part of an organization is exposed to the program while other parts are not. When selection to the program is not in the hands of the participants, and when the participants' level of need does not determine eligibility, the comparability of the groups is quite good. Unfortunately, as will be shown later, these favorable conditions are not often met. As mentioned in Chapter 9, the possibility that the meaning of the dependent variable has been affected by the program must be considered; if that is likely, a comparison between pretest and posttest would not be valid.

Problems in Selecting Appropriate Comparison Groups

Comparison groups are chosen according to the evaluation questions to be studied. A no-treatment group would be used when one seeks to learn whether the program has any effect. However, a no-treatment group would not be appropriate if stakeholders want to compare different ways of offering a service. In that case, groups being given different programs would be compared. If there is a suspicion that attention alone could affect an outcome, then the comparison group would be a placebo group, that is, a group that experiences a program not expected to affect the outcome variable.

Here are some other situations in which comparison groups are chosen in a way that introduces a bias in the evaluation. Let's imagine a school whose students tend to perform below the community averages. Suppose that on the basis of standard tests at the end of the third grade, the 20% of the students who are doing the worst are selected to be placed in a remedial program during the fourth grade. Suppose that the other 80% are treated as a comparison group and go through the regular fourth grade. After a remedial program, the children are retested. Because all children have been in school for the fourth grade, all should show some improvement; however, when the children are tested again, regression toward the mean will serve to raise the scores of the children who had scored the lowest. The scores of the upper 80% will suffer some regression toward the mean, thus lowering their scores relative to those of those students who had made up the lowest 20% on the first test. In this case, regression toward the mean will make the program look better than it is: regression contributes to calling the program a success. Analysis of covariance will not correct for the initial difference between the groups. With that difference of extreme high and extreme low scores, regression to the mean would actually affect the groups in different directions.

Now imagine another situation. The entire third grade is to get a novel reading curriculum. No comparison groups can be obtained from third graders so the evaluator decides to find fourth graders who read about as well as the third graders. Which fourth graders would that be? It can only be those fourth graders who are doing relatively less well compared to other fourth graders. Upon retesting, how will regression affect their scores? It will raise them. In this case, regression toward the mean would make the program look less successful than it is. This would be true because the third graders will get better scores because of an additional year of reading instruction, but there are two reasons why the lower scoring fourth graders will improve. These fourth graders are getting instruction and their test scores are regressing upward toward the mean of their group.

The emphasis on regression to the mean should not imply that this is the only weakness of nonequivalent control group designs. Regression was stressed because some stakeholders may mistakenly believe that selecting certain participants for inclusion in the evaluation can compensate for the effects of preexisting differences between the pretest scores of the groups. The moral is clear: the nonequivalent control group design is especially sensitive to regression effects when groups are systematically different on some dimensions. Because a systematic preprogram difference between treatment and comparison groups is likely to be found in compensatory education evaluations, the nonequivalent control group design can be applied only with caution.

Here is a third situation for which comparison group selection could be difficult. Suppose that a voluntary program is to be evaluated. Imagine parents seeking out a special program for their children. The natural comparison group would be children of the same age and from the same schools whose parents have not

sought the special program. It is possible that those children whose parents choose to place them in a special program are maturing faster than those in the comparison group. Parents who seek out special programs for their children may also be devoting more attention to their children at home than are parents who do not seek out special programs.

Evaluators seeking a solution to the problem of pretest differences between the program and the comparison groups have matched people in the program group with those not in the program. People are chosen to form a comparison group because they resemble the program participants on income level, objective test score, rated adjustment, locality of residence, or other criteria. Even this level of effort at matching will not overcome the fact that the groups are different in critical ways that will bias the interpretation of the evaluation.

There are many other reasons why existing groups may differ from each other. One might compare neighboring classrooms in an attempt to learn if a novel teaching method used in one classroom is more effective. Although the children in the two classrooms seem equivalent, it might be learned later that the teacher in the comparison (i.e., nontreatment) classroom had been using methods that are similar to those used in the program. Or, when patients of two physicians are used in an evaluation of a patient education brochure, it may turn out that the physician handing out the brochures has systematically encouraged her patients to read about their illnesses and treatments. Thus, her patients are more likely to read such material than are the patients of other physicians. This preexisting difference between patient groups may make the patient education program appear more effective than would have been the case if a different physician had been asked to distribute the material.

In sum, although differences in groups at the time of the pretest can lead to over- or underestimates of the effects of programs, the nonequivalent comparison group design is quite powerful and useful when similar groups are available to form treatment and comparison groups, as long as its limits are recognized and addressed.

REGRESSION DISCONTINUITY DESIGN

There is one situation in which a comparison between nonequivalent groups can be made in a manner that is much more powerful than the methods presented so far. When eligibility for a service is based on a continuous variable, such as income, achievement, or level of disability, it may be possible to use the regression discontinuity design (Trochim, 1984, 1986). The strength of this approach has been recognized by the What Works Clearinghouse as one of three designs that are eligible to receive the highest rating "meets standard without reservation" (Louie, Rhoads, & Mark, 2016), and some have suggested that its prospective use is particularly valuable (Klerman, Olsho, & Bartlett, 2015).

Suppose, for example, that 300 students are tested for reading achievement in September. Those scoring the lowest are defined as those most in need of extra assistance. If the program has facilities for 100 students, it seems reasonable and fair to take the 100 with the lowest scores into the program. This strategy is simply a special case of the nonequivalent control group design; however, unlike other ways of choosing a comparison group, the evaluator has a quantitative measure of how the two groups were formed.

If all 300 are retested at the end of the school year, what would be expected? Note that we would not expect the 100 to outperform the 200 regular class students. If the program were effective, we would instead expect that the treated children would have gained more than they would have had they stayed in their regular classrooms. The regression discontinuity design enables the evaluator to measure such effects.

Figure 10.4 illustrates the key concepts. The scatterplot on the left shows how we would expect a pretest and a posttest to be related; there is a strong, linear, positive correlation. The pretest and posttest do not have to be identical; they simply have to measure the same construct—reading skills, mastery of chemistry, and parenting skills. If those who are low on the pretest received an effective program that added two points to what the participants would have had on the posttest, the scatterplot would look like the plot on the right side of Figure 10.4. To measure the effect of the program requires the use of multiple regression, as explained in Box 10.1. The scatterplots in the box show why evaluators need statistical analyses. Without the regression analysis, it would have been impossible to notice the sizable effect of the program. The regression analysis does detect it, and plotting the regression line displays the discontinuity very clearly. If the special program

were equivalent to the regular class, one would expect a smooth relationship between pretest and posttest. If the program were actually (heaven forbid!) worse than the regular classes, we would expect program students to score lower than expected on the basis of the pretest. In such a situation the discontinuity would be the reverse of that in Figure 10.4 and Box 10.1.

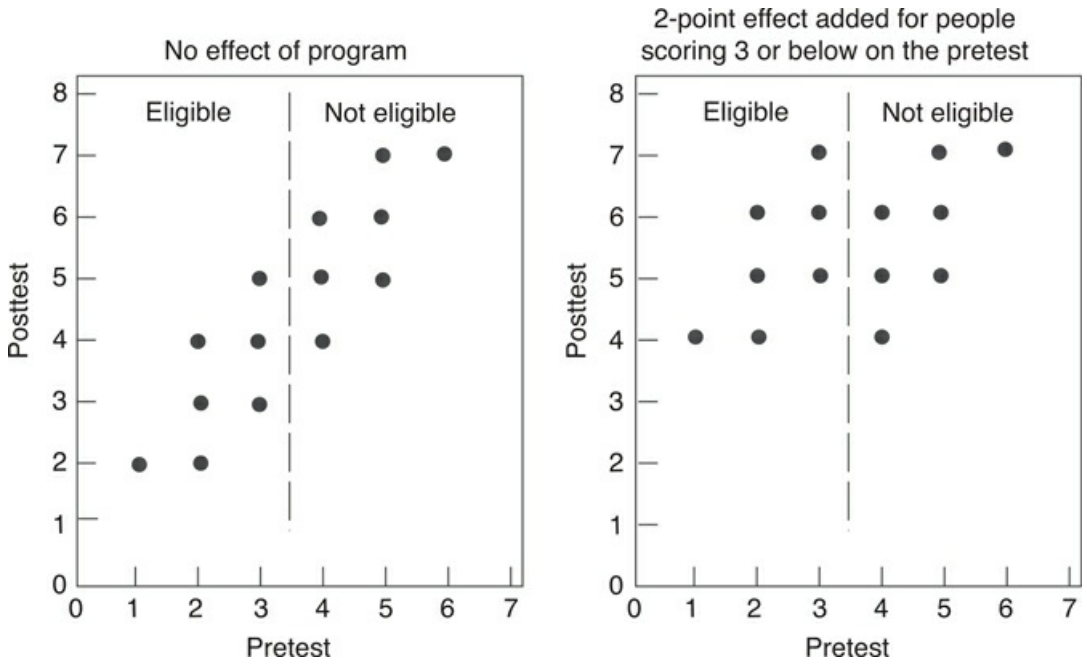


Figure 10.4 The relationship between pretest and posttest scores when people scoring three or below are eligible for special assistance to improve posttest performance. The graph on the left illustrates the case when the assistance was not effective; the graph on the right illustrates how the data would appear if the program enabled the eligible people to perform two points better than they would have without special assistance.

Box 10.1: How to Perform and Interpret a Regression Discontinuity Analysis

If you have studied multiple regression, carrying out a simple regression discontinuity analysis is very easy. First, one needs pretest and posttest information on a group of people. Only the people on the low (or high) end of the pretest are exposed to the program. We must have two observations for all people regardless of their scores. Suppose that students with less than 60% on the pretest, a placement test, were eligible for remedial help prior to taking a regular course. To evaluate the effectiveness of the remediation, we can use a regression discontinuity design. The posttest, the final in the regular course, is the dependent variable (the variable to be predicted). The placement test is one independent variable, X_1, and getting remedial help or not is the second independent variable, X_2. The value of X_2 for the 100 students getting help is 2; the value of X_2 for the 200 students not getting help is 1.

We have every reason to expect that the placement test will be a strong predictor of the final in the course. If the remediation was helpful in improving scores on the final, then X_2 will also be a strong predictor of the final.

The figure below (left side) shows a scatterplot in which program participation led to improved posttest scores. The regression line is shown in the figure; the program effect creates the discontinuity seen in the sudden drop at the dividing line between the people with eligible scores (60 or lower) and those not eligible (above 60). The regression equation supports the conclusion that participation in the program led to higher scores on the final.

Predicted final = 21.25 + 0.57 × (placement test score) + 7.67 X_2 where X_2 equals 1 for those not eligible and 2 for those eligible.

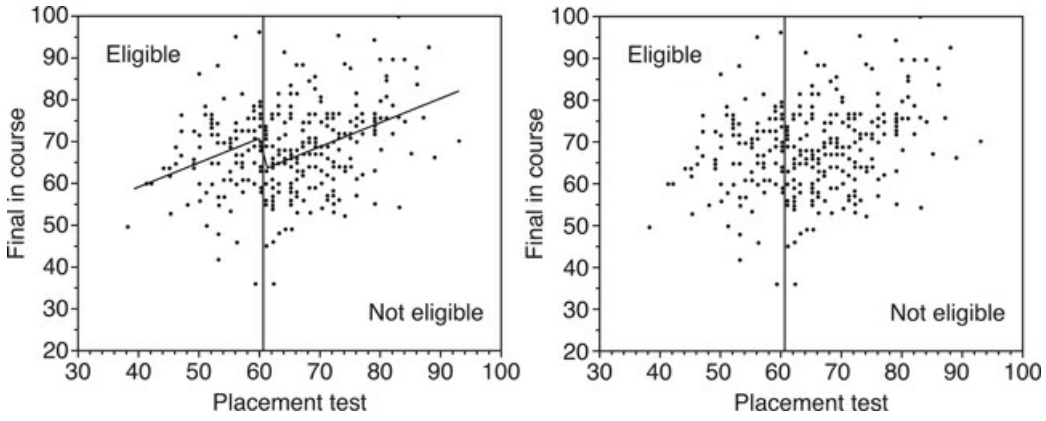
The relationship of being in the program (or not) with the final in the course was statistically significant ($t(297) = 4.027$, $p < 0.0005$). The effect of the program was +7.67 points. The effect size of the program is

$$g = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 4.037\sqrt{\frac{1}{100} + \frac{1}{200}} = 0.493$$

This effect size indicates that the remediation produced a moderately large effect according to the findings of Lipsey and Wilson (1993) mentioned in Chapter 9.

Additional details on using regression discontinuity analyses in more complex situations are given by Trochim (1984, 1990) and his

colleagues (Reichardt, Trochim, & Cappelleri, 1995). Although the examples of regression discontinuity designs are usually presented in the context of compensatory programs for those who are in danger of falling behind, they are equally relevant to evaluate programs for the gifted and talented (Marsh, 1998). Note the scatterplot on the right below. This is the same scatterplot as on the left, but without the regression line. It is not easy to detect the effect of remediation without the help of a regression discontinuity analysis.



OBSERVING OTHER DEPENDENT VARIABLES

At this point the methods described in Chapter 9 have been expanded to include observations of the program participants at many times and to include observations of groups other than the program participants. It is possible to increase the validity of interpretations by observing additional dependent variables that are not expected to be changed by the program, or, at most, changed only marginally; this design has been called the *control construct design* (McKillip, 1992; McKillip & Baldwin, 1990).

The added dependent measures must be similar to the major dependent variable without being strongly influenced by the program. If children do read better after a special program, it might be expected that they would do better on a social sciences test than they did before the reading program; consequently, a social sciences test would not be a good control construct measure for a reading program. However, a test of mathematical skills requiring only minimal reading levels might be an excellent additional dependent variable. Patients in a health education program concentrating on avoiding heart disease should show increased knowledge about heart functioning, cholesterol, and related topics. They would not be expected to show greater knowledge of osteoporosis. When it is possible to find additional dependent measures that (1) would be affected by the same threats to internal validity as the outcome measure, but (2) would not be affected by the program, this design is a powerful one. This design was used in evaluating the Medicaid-required second opinion on surgical procedures by comparing the percentage reductions for ten procedures requiring a second opinion with the reductions for nine procedures not requiring second opinions (Tyson, 1985). The effects of a media-based alcohol education campaign on a university campus were evaluated using control constructs that were alcohol related, but were not the subject of the campaign (McKillip, Lockhart, Eckert, & Phillips, 1985).

COMBINING DESIGNS TO INCREASE INTERNAL VALIDITY

Time-Series and Nonequivalent Control Groups

The most interpretable quasi-experimental designs are those that combine the approaches described above. If a group similar to the program participants can be observed, the simple interrupted time-series design is strengthened considerably. Furthermore, if the effect of a program can be replicated, a very credible argument for the effectiveness of a program can be made. Figure 10.5 is a hypothetical example of such a finding.

The problem of physical abuse, neglect, and financial abuse of the elderly has come to the attention of the public (National Center on Elder Abuse, 2005). If a community wanted its police department to make officers more aware of this problem, a training program might be developed. The program would train officers to better detect situations of abuse of elderly people. Although no one wants to see greater number of arrests for any crime, it is important to detect crimes that do occur. If there is a problem with elder abuse, an evaluator would expect to observe that arrests would increase after the program. Suppose that arrests for elder abuse are

plotted per month for 6 months before the program and 12 months afterwards. This is called an interrupted time-series design. It is likely that arrests will vary a great deal from month to month because such abuse is indeed rare, but the number of arrests is expected to increase. An alternative interpretation of an increase might be that the general level of awareness of the problem has become widespread and the increase is really due to that increasing general awareness, not the program. That alternative interpretation could be examined much more validly if arrests from other communities that did not initiate similar programs were also available for the same time periods. All communities might be affected by the greater awareness of the problem, so it would not be surprising if the number of arrests were increasing; however, in the community whose police were especially trained, one would expect a larger increase. Having the comparison communities permits an evaluator to detect change greater than historical trends perhaps seen in other communities, thus demonstrating the effectiveness of the training. Hypothetical graphs are shown for the first community in Figure 10.5. If a program was offered over a weekend at the end of month 6, an increase for months 7 through 12 would be expected, but an increase is not expected for the second community.

An especially strong evaluation design in this context would be obtained if the second community instituted the same training at a later time. In that case, an evaluation would include a replication of the interrupted time series. If the training had been effective, a time-series graph containing the trend lines can be smoothed for the first community and for the second community as in Figure 10.5. If the training in the second community had occurred between months 12 and 13, the hypothetical graph shows evidence for a replication of the effect of increasing police officers' sensitivity to the issue. The pattern in Figure 10.5 would be strong evidence for the effectiveness of police training. By the way, there might be a temptation to use a two-way analysis of variance to analyze the arrests per month using the communities as one independent variable and three six-month periods as the second; however, these data would not satisfy the assumption that each observation is independent of the others. Instead many methodologists (e.g., Riecken & Boruch, 1974) have argued that tests of statistical significance are less important than a qualitative understanding of the various threats to the validity of causal conclusions about the impact of an intervention. Figure 10.5 illustrates an *interrupted time series with switching replications* (Shadish, Cook, & Campbell, 2002). If the observations fit the pattern in the figure, little in the way of statistical analysis needs to be done.

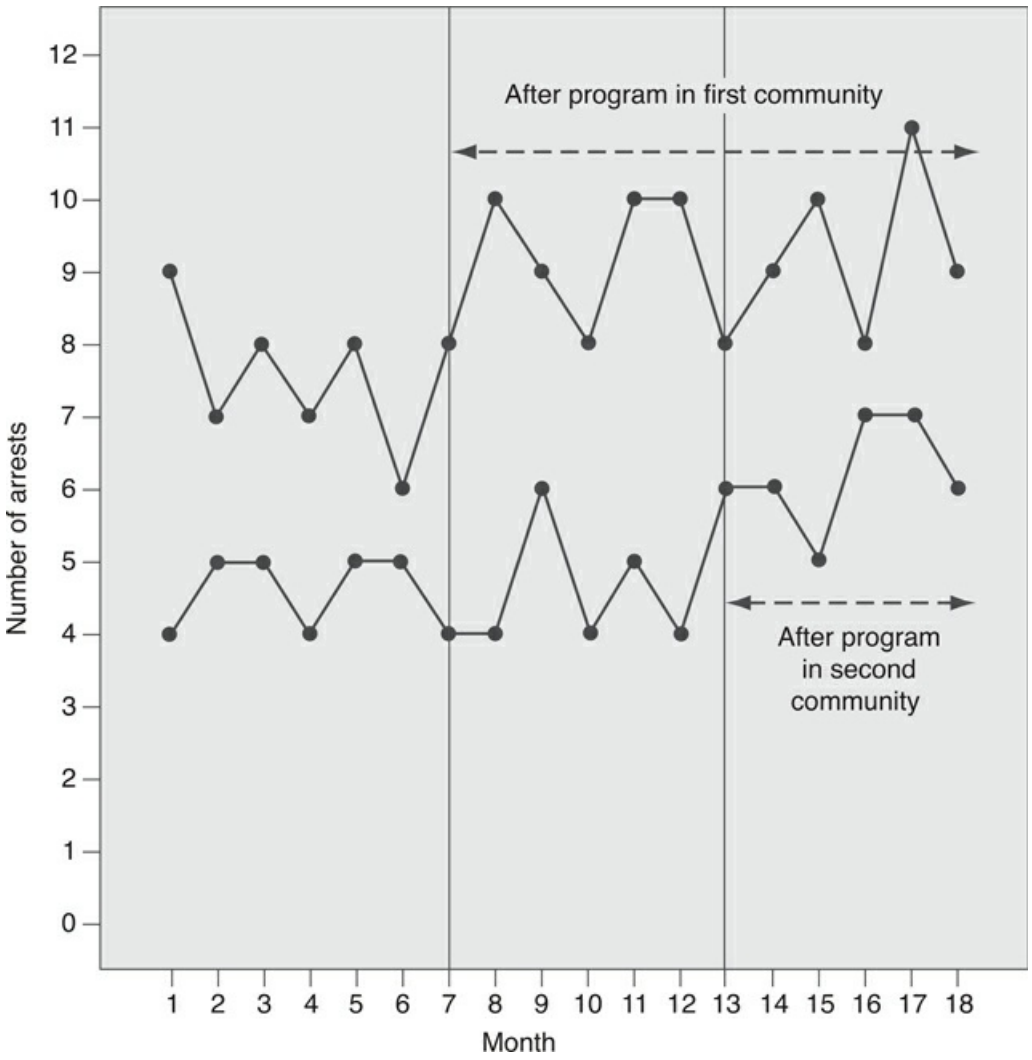


Figure 10.5 Hypothetical results of an interrupted time-series design with switching replications.

Selective Control Design

By understanding the context of a program, evaluators may be able to identify the threats to internal validity that are most likely to affect an evaluation. Evaluators may then choose comparison groups to control for specific threats to internal validity so that the most plausible alternative interpretations are eliminated. When the appropriate nonequivalent control groups are available, the *selective control design* can be a powerful approach (Cordray, 1986).

Here is a simple example of planning an evaluation based on a selective control design. The dean of a college of arts and sciences requested an evaluation of a junior year abroad program offered by the college. The report was needed for a board of trustees meeting that was to occur in six weeks. Clearly, there was no opportunity for a time-series design. (As mentioned earlier, it is not uncommon for evaluators to work under a tight deadline.) The evaluation was planned around the obvious comparison—college seniors who had studied abroad the previous year versus seniors who had remained at the parent campus. In making this comparison, self-selection as a threat to internal validity cannot be ignored. Students who study abroad for a year are different from those who remain at home. One could estimate the preexisting differences between these two groups by testing individuals before they went abroad, but a pretest–posttest design was impossible due to the short time allotted for the study. Because the decision to study abroad requires much planning, students sign up early for the program. By comparing seniors who have been abroad with sophomores arranging to go abroad, the self-selection threat could be addressed because both groups have selected themselves into the program. However, now a second threat to internal validity becomes a problem: sophomores are less mature than seniors. Two years of age may not be that critical for middle-aged and elderly people, but two years may

make quite a difference for an evaluation of a program for college students. By adding one more group, sophomores who did not intend to study abroad, the major threats to the internal validity of this evaluation were controlled.

The design is summarized in Figure 10.6. If self-selection were related to high scores on dependent measures, one would expect the groups in the upper row to have higher scores. If maturation led to higher scores, those in the right-hand column should have higher scores. If the program, junior year abroad, had an impact, then the upper right-hand group, the self-selected seniors who studied abroad, should have especially high scores as illustrated. If students studying abroad developed as hoped, an interaction would have been found in a two by two analysis of variance. Unfortunately, the sophomores planning to go abroad could not be surveyed in time for the trustees' meeting; consequently, the means in the figure are hypothetical.

		Maturation Level	
		Sophomores	Seniors
Self-Selection	Students who will study or have studied abroad	50	65 (Treatment group)
	Students who have not studied or do not plan to study abroad	40	45

Figure 10.6 Hypothetical mean scores on a measure of international understanding (high numbers indicating more international understanding) for the students in the junior year abroad evaluation. A program effect is indicated by these means.

The design did not deal with some threats to internal validity because they were not plausible in this setting. For example, instrumentation refers to possible changes in measures of outcome because the observer becomes more vigilant or experienced during the course of an evaluation. The dependent variables were self-report and every student would have taken the attitude test just once; thus, neither instrumentation nor testing effects could affect the findings.

Case Study 10 makes a couple of important contributions to this text. First, it represents an evaluation based on the focus of this chapter—illustrating both what can be learned from quasi-experimental designs as well as the limits to such evaluations. Second, it also provides the foundation for seeing how evaluations develop over time. This account was written in late 2017, while the evaluation was still in progress. The eResources for Chapter 10 will include updates at least once per year on this evaluation, showing how this project continues to develop.

CASE STUDY 10

Learning How to Evaluate—Using PCOMS

As noted in the text, this case study describes an evaluation that is being initiated as this edition is being written, so the results are not yet in place. Regular updates will be posted in the online resources for this book, so students can see how the project is actually developed. The project involves evaluating "the effectiveness of implementing client feedback training on counselor self-efficacy, therapist development, and client outcomes."

Specifically, Dr. Norah Chapman oversaw having students in practicum classes in the doctoral clinical psychology program begin using two very brief measures every session with their clients as a program initiative. The measures assessing client outcome and therapeutic alliance were expected to lead to positive outcomes for clients and therapists as well as to document the doctoral program's training in greater detail.

Meeting Needs

The new use of the measures addressed several needs at once. Clients need high-quality treatment, students need high-quality training, and doctoral programs need evidence related to both as a part of their accreditation process. The evaluation proposal addressed these varied needs and cited research showing that the use of client feedback has been found to support these desired outcomes. The evaluator

noted in particular that the measures were intended to allow the program to assess both positive and negative effects of student practicum work—catching any problems before they occur or worsen, but also providing evidence of successes.

Implementation

The evaluation proposal followed the initial implementation of the use of the measures by practicum students over the previous two years. Although program faculty had already considered various aspects of the use of the measures, the evaluation was intended to provide a more systematic assessment of the entire process of implementation.

Stakeholders

Four distinct sets of stakeholders are noted: clients, practicum students providing psychological services, the doctoral training program, and APA—the accrediting agency.

Side Effects

Although the program evaluation has not yet been completed, some negative effects observed already were some "glitches" in the data management system for recording feedback data and the additional time required to address those problems. Others will be noted over time.

Improvement Focus

The expectation is that evaluation will allow the program to learn more about the factors most relevant to issues of training. The upfront commitment of the program faculty and administration to the evaluation is an encouraging sign of the potential for substantial use of evaluation insights.

Outcomes

As noted, multiple outcomes are expected—better client functioning, better student training, and better program documentation. Although it is hoped that all these improve, there may be important differences among them.

Nuances

One expected mechanism is the way that practicum sites do or do not support the active use of the measures. Although other potential mechanisms might be imagined, the limits of time available during practicum therapy sessions mean that only very limited exploration of these mechanisms is possible.

Summary

The quasi-experimental design of this evaluation, although not as powerful as a full experiment, still shows important strengths beyond the one-group designs common in many evaluations. Among other points, findings for students in their second and third year of practicum placements provide an important comparison with those in their first year, showing any developmental effects of increased learning or simply greater amounts of training experiences. Even more than just groups with different amounts of training and experience, as new cohorts enter the program, the same students will have data for multiple years of work with clients, so differences across years for those students will be stronger evidence of changes. Be sure to check the eResources for recent updates.

Source: N. Chapman (personal communication, November 15, 2017).

EVALUATOR PROFILE 10

Norah Chapman, PhD—A Two-Way Relationship

Dr. Chapman's experience in her doctoral program in Counseling Psychology at the University of Kentucky provided an important base for her interest in using the Partners for Change Outcomes Management System (PCOMS) (Duncan, 2012) with students in Spalding University's PsyD program. Her own program's Scientist–Practitioner model clearly indicated that science informs clinical practice, and she saw good reason for that process in her coursework. But when she started her practicum experiences, she saw the corresponding reality that clinical practice informs science, especially as she saw the value of naturalistic research rather than only basic scientific studies.

As she used PCOMS in her own training, it was a natural tool to consider when she began her career. When she was hired at Spalding on a one-year contract, she saw the opportunity for its use, but several aspects of her situation meant that using PCOMS would best wait until another time. Fortunately, events shifted her perspective—the next year, she was hired in a tenure-track position, and the prospect of a long-term involvement meant that putting time and effort into helping students learn how to use the system was a good investment for them and for her. Among other things, PCOMS fit the needs of Spalding's students. It is an evidence-based practice and Spalding's School of Professional Psychology is committed to such practices. Further, there were multiple layers of application—in addition to

helping therapists evaluate clients' progress as well as helping clients understand and track their own therapy, it supports broader evaluations of the program.

On a broad level, it had the potential to document the overall effects of all practicum students' work with clients, rather than just assuming that the large amount of clinical work was producing good results. Progress of cohorts could also be assessed—experienced practicum students should be facilitating at least as much positive change as practicum students early in their training. The specific data could support analyses that previously could only be imagined, such as specifying expected outcomes by practicum setting, especially combining similar populations and comparing them with others. For example, it is commonly assumed that chronically mentally ill patients generally show modest gains in contrast to first-time clients who demonstrate more dramatic improvements, but rather than relying only on assumptions, hard data from the measures could demonstrate when the assumptions were accurate and when they were not supported.

Norah oversaw the implementation of Spalding practicum students using PCOMS. Among other points, she notes how it supports her own commitment to building social justice: "It gives clients a voice in a system that generally disenfranchises those with mental illness." As noted in the case study, there will be regular updates in the eResources as this project continues to unfold. There will also be the opportunity for Dr. Chapman to reflect on her experience with this developing project.

Source: N. Chapman (personal communication, December 5, 2017).

SUMMARY AND PREVIEW

The use of quasi-experimental designs in program evaluation is widespread. These approaches often minimize disruption yet permit gathering sufficient information to isolate the probable effects of innovative programs. However, using quasi-experimental designs must be thoughtful; evaluators need to consider which specific threats to internal validity must be accounted for in each evaluation. Anticipating these threats can cause headaches for evaluators: there is no standard approach to planning quasi-experimental evaluations. Sometimes the preferred statistical methods to use with these designs are controversial.

When conditions permit a controlled experimental evaluation, evaluators often take the opportunity so that few threats to internal validity are able to bias the findings. When an intervention is expensive, affects many people, or might lead to serious side effects, it is especially desirable to conduct an experimental design. Chapter 11 describes the value and some of the limitations of evaluating programs with experiments.

STUDY QUESTIONS

1. Explain which threats to internal validity are not covered by an evaluation of a parole program in which male parolees who are tested before and after the program are compared to a sample of men living in the same neighborhoods as the parolees and tested at the same times.
2. Assume that reading levels in a school district had been steadily declining over a period of years and that a new reading program was implemented to counter the trend. After two years, reading levels declined again, but the superintendent announced that the reading program was a success because reading levels would have declined even more without the program. What would be necessary to evaluate the validity of such an interpretation?
3. Suppose that a new treatment were offered to patients of a group medical practice in an upper-middle-class suburb. As a comparison group, similarly aged patients with the same diagnoses were found in a county hospital providing free (Medicaid) coverage to poor families. Evaluate this nonequivalent control group evaluation design.
4. A problem with quasi-experimental evaluations is that often a particular set of results can be interpreted, but a different set cannot. For example, assume that a service to a disadvantaged group is supposed to lead to higher scores on some variable (see question 3). If the post-program level of this disadvantaged group is higher than that observed in a middle-class comparison group, the interpretation of the evaluation is relatively easy. However, if the postprogram level of the initially disadvantaged group improves, but is still lower than that of the middle-class comparison group, then there are a number of possible interpretations. Explain these statements.
5. If you did not plot and smooth the time series given in this chapter, do so now and carry out the smoothing process.

ADDITIONAL RESOURCE

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Boston, MA: Houghton Mifflin.

This is the same additional resource as presented in Chapter 9 because it continues to be the standard research design resource for program evaluators. Chapters 4 through 7 provide more details on the quasi-experimental designs described in this chapter.

11 Using Experiments to Evaluate Programs

Even after carefully designed quasi-experimental evaluations are completed, concerns may linger that some effect other than the program caused the change in the participants or masked a real program effect. A way to minimize these concerns is to adopt an evaluation design strategy that controls for threats to internal validity better than quasi-experiments. The easiest way to do this is to use an experiment based on observations of people randomly assigned to participate in the program or to some alternative experience. This chapter reviews the need for experimentation as a tool in program evaluation, describes the most opportune times to implement an experiment, and makes practical suggestions for dealing with the problems of conducting experimental research in organizational settings.

EXPERIMENTS IN PROGRAM EVALUATION

Benefits of Experiments

Experiments are not always possible in evaluations. Quasi-experimental designs may be the only option when it would be unethical to assign people to experience or not experience a program or when organizational conditions are such that random assignment is not practical. In such settings evaluators utilize existing intact groups. Such groups develop for reasons other than the need for an unambiguously interpreted evaluation, so consequently, it is very likely that the groups will not be equivalent before a program begins, and the lower level of confidence in the meaning of the results are understood as an unfortunate necessity.

Several methods have been used to make groups more comparable—matching, finding similar groups, or using statistical adjustments. In some situations these approaches serve fairly well, but often nothing suffices to eliminate fully the threats to internal validity. Experiments avoid these problems to a substantial degree.

Experimental Designs

The simplest experimental design involves randomly assigning people either to a program group or to a control group and then contrasting the scores of the people in the groups on observations made of outcomes after the program. If the groups are formed randomly, there is less need for pretests because the groups are less likely to be substantially different before the experiment begins. Consider some of the threats to internal validity that an experiment controls: (1) the groups should not differ in rate or level of maturation, (2) people cannot select themselves into one or the other group, (3) although *individuals* begin with different degrees of need, the *groups* do not begin the experiment at different levels on the outcome variable, (4) all people experience the same amount of testing, and (5) all groups represent the same population so the groups cannot be regressing toward different means. In other words, if everything goes right, the postprogram group means can be validly contrasted using standard statistical techniques. (Of course, random assignment only means that differences were not deliberately created or present for other reasons such as self-selection. Some differences can occur even with random assignment.) Having more than two groups simply requires that analyses of variance be used instead of *t*-tests.

If possible, it is always advisable to measure program participants prior to the beginning of a program. If people have been randomly assigned to groups, the means of the groups will probably not be systematically different before the program begins, but measuring allows testing of this assumption. A measure of the desired outcome made before program participation is expected to be positively correlated with a measure of the outcome after participation for both program and control groups. Having a preprogram measurement that

is correlated with the outcome variable permits the use of analysis of covariance (ANCOVA), a statistical test that may reduce the likelihood of Type II errors in comparison with an independent groups *t*-test or ANOVA.

ANOVAs, ANCOVAs, and *t*-tests are just some of the analyses that may be needed for particular experiments. See the Chapter 11 folder in the eResources for a set of materials about a range of analyses, including a guide to the appropriate analysis, displays of results with comments on the specific elements such as probability, a guide to interpretation of results, and links to other online resources. Although inferential analyses like these do not solve all issues of evaluation, many evaluators find that statistics can be an important obstacle, so good resources can be very helpful.

Using pretests provides evaluators with another advantage: they can help in describing findings to stakeholders. If it is possible to measure the variable that will be used as the outcome both before and after the program, one can provide an estimate of improvement during participation in the program; many stakeholders will value such information. There is another advantage to using pretests when one can be sure that their use does not create a problem by sensitizing participants to the outcome measure (a threat to internal validity—testing). Adding pretests to the basic experimental design assures evaluators that if problems were to develop during the process of randomization, some analytic options are still available. Unfortunately, sometimes something does go wrong, such as when there are different rates of dropouts from the experimental and control groups. If so, the data can be treated as a pretest–posttest nonequivalent control group design.

Another refinement in forming program groups randomly is to place similar people into groups before randomization. Perhaps people might be grouped or even matched on the basis of age range, skill level, severity of disease, or previous experience in similar programs. Then people are assigned randomly to participate or not participate in the program from all groups; that would guarantee that people representing all levels of the relevant variable would be in both program and control groups. The advantages of this refinement include having the option of analyzing the data to examine the relationship of the variable used to match the people. It may be that older participants are more strongly affected by a program than younger participants. If that were true, program directors would want to know.

OBJECTIONS TO EXPERIMENTATION

Although experiments have clear advantages in making evaluations interpretable, many people object to experimentation in service settings. It is important for evaluators to be familiar with the most common objections, lest they encounter them for the first time while they are planning an evaluation or, even worse, after random assignment to groups has already begun.

Don't Experiment on Me!

Some people are not eager to participate in an experiment perhaps because experimentation has sometimes been associated with the "mad scientist" of science fiction. Moreover, the popular media do not use the word *experiment* in the careful way experimental design textbooks do. Consequently, it is not surprising when people not trained in social science methods equate experiments to evaluate the effects of social programs or medical treatments with undisciplined attempts to learn about behavior. Evaluators planning an experiment will be careful to treat the question of experimentation with great respect. The advantages of experiments should be stressed and the fact that the innovative program has been carefully planned by competent personnel should be made clear to possible volunteers and committees overseeing the research. Further, the similarity of the treatment to other options can be noted. Guarantees of confidentiality should be made clear.

We Already Know What Is Best

Experiments are sometimes carried out on treatments that many professionals have come to accept. Particularly tragic are ineffective medical interventions that have negative side effects. A classic example still illustrates this important point. Nonexperimental observations had led physicians and nurses to believe that pure oxygen was essential for the survival for premature infants; however, at first, it was not known that using pure oxygen would also leave some infants blind. Once this terrible side effect was suspected, it became

important to learn whether it was necessary to use pure oxygen with premature infants. During research on the value of oxygen, some nurses who thought they knew best switched the oxygen on at night in order to be sure that the babies in the regular air group got some pure oxygen. Once controlled research was complete, it was learned that the administration of pure oxygen was indeed responsible for abnormal growth in the developing retinas of infants. Only later was it learned that air with a higher concentration of oxygen (but not pure oxygen) could be administered to increase the survival rate of premature infants, yet with little chance of hurting the development of their retinas (Silverman, 1977). Inadequate evaluations of the effects of drugs administered to coronary patients to stabilize irregular heartbeats resulted in the premature deaths of tens of thousands of patients between 1985 and 1989 (Moore, 1995). Even bed rest, an innocuous-sounding and widely recommended treatment, has negative side effects; evaluations, when done, have shown negative results (Allen, Glasziou, & Del Mar, 1999).

I Know What Is Best for My Client

Some service providers are convinced that they know what is best for specific clients, students, or patients. Consequently, they are not eager to commit themselves to follow the treatment dictated by a random assignment procedure. Similarly, people are hesitant to become involved in experimental evaluations of services because they believe that human service providers can choose the service that is best for meeting their needs. In fact, the reasons for the choice of treatments provided to many people are quite unknown. Different counselors follow different theories of psychotherapy, physicians do not follow set treatment protocols (Eddy, 1990), different cities and states set up different welfare policies, and judges do not assign identical sentences to people convicted of the same crimes. Although there may be quite valid reasons for choosing different treatments, an individual provider does not make the same choices even when faced with clients with the same problems and different caregivers are not consistent with each other (Dawes, 1994).

It takes considerable tact to explain to service providers that their beliefs may not be always valid. Providing examples of accepted practices that were later invalidated by controlled studies may help providers see the possibility of misperceptions and the advantages of unambiguous documentation of service effectiveness. Evaluators often suggest that society would benefit by treating new human service programs as if they were experiments—treating failures as an indication of our incomplete understanding of people and organizations, not as an indication of low motivation or incompetence of service providers (Deming, 1986). Unfortunately, some human service providers act as though a negative evaluation indicates that they are at fault rather than their theories or techniques; therefore, they develop rationalizations whenever particular individuals do not seem to benefit from a service. By explaining away all failures, program personnel miss the opportunity for growth that could occur if they improve their practices on the basis of feedback. At these times evaluators feel the full force of the conflict between their skeptical role toward service programs and the providers' confidence in the effectiveness of the services they give.

Experiments Are Just Too Much Trouble

Anyone who has conducted an evaluation in an organization knows that a commitment of time and effort is needed (see Case Study 11). Experiments in service delivery settings require a lot of nurturing to plan, implement, and monitor (Boruch, 1997; Dennis, 1994). Evaluators should not agree to conduct an experimental evaluation if they do not have access to the resources needed to carry it through properly; however, it may take only a little additional effort to carry out an experimental evaluation compared to a noncontrolled study. If hospital patients are assigned to different general medical hospital units on the basis of bed availability, the patients in these units form the equivalent of randomly assigned groups. Carey (1979) was able to conduct an evaluation of alternative approaches to the organization of hospital nursing because different forms of nursing were implemented on the two general medical floors of a hospital. On one floor nurses followed the widely used team approach in which different nurses take specific responsibilities for many patients (e.g., administer medications), while on the other floor nurses followed the primary nursing model in which a single nurse is responsible for all of the care needed by four patients. In this case, conducting an experiment did not add any burden to the hospital beyond the effort of data collection. By the way, little difference between the two approaches was found—a finding replicated by others (McPhail, Pikula, Roberts,

Browne, & Harper, 1990). A research center was based on four parallel medical care provider groups making up the center. Patients were randomly assigned to one of the groups (Neuhauser, 1991). Contemplated changes in care techniques were evaluated in one or two of the firms before the successful changes were implemented in the entire health care center.

What is often overlooked in considering the cost of an experimental evaluation is the cost of *not* doing an experiment. Several examples were suggested above of situations in which medical care that was believed to be beneficial actually harmed patients. Reviews of research on several medical treatments that at one time had widespread support revealed that they had never been evaluated adequately. The classic comment of Gilbert, Light, and Mosteller (1975) that the alternative to experimenting on people is "fooling around" with people illustrates this important point. Seldom do service providers inform their patients, clients, or trainees that the service or care offered has not been fully evaluated, even though that is often true.

THE MOST DESIRABLE TIMES TO CONDUCT EXPERIMENTS

A theme of this text is that evaluation should be built into the daily operation of all organizations. Usually such evaluation practices will not be experiments but rather the administrative-monitoring and performance-monitoring approaches described in Chapters 7 and 9. The experimental model is not the all-purpose program evaluation methodology (Lipsey, Crosse, Dunkle, Pollard, & Stobart, 1985). There are, however, circumstances that are best served by an understanding of cause-and-effect relationships that can be best studied with experiments (Cook, 2006), and the demand for evaluations that address causal relationships is generally growing (Rog, 2015).

When a New Program Is Introduced

When a program is new, there are several conditions likely to increase the interest of staff and management in an experimental evaluation. First, when people are to be randomly assigned to receive the current, standard form of service in order to form a control group, it is hard to argue that they will be treated unfairly.

Second, a new program may take some time to implement, thus providing time for an evaluation. In large organizations, many new programs go into effect in phases; evaluators can use such delays to conduct evaluations. The implementation of training, the delivery of new equipment, or a reorganization seldom takes place in all divisions of a company at one time. It might be possible to schedule the locations of the introductions randomly. Then, as the program is begun in different divisions or locations, several groups will be receiving the program at the same time, but in different locations. Ideally, the variables that are expected to improve will change as the program is implemented. If similar impacts are observed at different times, history, maturation, and regression toward the mean are not plausible alternate explanations for change. In order for administrators to schedule implementation in ways that permit an evaluation, evaluators must have input into the planning committee and administrators must value information on the effectiveness of their decisions.

Third, when there is a need to document the program's success, experimentation may be desired. This documentation may be for internal or external purposes. In the case of demonstration projects funded by grants, funding agencies want evidence that a program might work elsewhere. A fourth reason why new programs are more likely to be evaluated is because there may be less anxiety over the possibility of discovering disappointing outcomes than when evaluating a program that became widely accepted before being rigorously evaluated. Regardless of the type of organization, evaluators working with new programs need to encourage a spirit of exploration among the staff because even well-planned innovations often do not work as well as hoped (Dial, 1994; Light & Pillemer, 1984).

When Stakes Are High

When a program is expensive, when large numbers of people are involved, or when the behaviors to be changed are important, controlling all possible threats to internal validity is desirable. Welfare policies, the organization of school districts, safety regulations for automobiles, and efficient provision of medical care are issues that have a great impact on many people. Proponents of programs to relieve the effects of social problems or to reduce the risk of injury and disease advocate many new services or policies. Since no society

can carry out all that might be done, selecting the best use of funds is wise. As discussions of policy failure and negative side effects of programs show, resources have been wasted when support was given for policies that had not been shown to be effective and efficient or that may even have had the reverse of the impact desired (see Howard, 1994; Sieber, 1981).

Sometimes when the stakes are high, proponents want to avoid taking the time for experimental evaluation because there are people clearly in need who might benefit from the intervention. Evaluators know that sometimes such enthusiasm is misdirected; however, it is important to encourage evaluation without appearing to be a cynical naysayer. One way to carry out an evaluation in such a context is to examine the criteria used to determine need. If there are people whose levels of need fall between those without needs and those clearly in need, perhaps an experiment can be conducted with those people (Crano & Brewer, 1986). This method of forming experimental groups was used to assign first-year students with marginal writing skills to either regular English sections or remedial English sections (Aiken, West, Schwalm, Carroll, & Hsiung, 1998). Those with definitely deficient skills were provided with the remedial program while the program was evaluated with those showing less dire levels of need.

When There Is Controversy about Program Effectiveness

No reader of this book would be surprised to be told that stakeholders often disagree about whether a program is likely to be effective or even what defines effectiveness. Recall the controversy over workfare programs highlighted in Figure 5.2. In addition, note the issues raised in Case Study 11.

When Policy Change Is Desired

Program evaluations based on experimental methods are sometimes implemented when there is widespread dissatisfaction with current policies, but no consensus on what changes to make. A desire to change welfare policies accompanied by uncertainty about how possible policies would affect people led to experiments with alternative welfare policies (Rogers-Dillon, 2004). The need to prevent, or at least to reduce, the rate of school failure of disadvantaged rural and urban children led to a series of experiments with Head Start, a preschool program aimed at improving cognitive skills, health, and nutrition (Datta, 1976).

When Demand Is High

An experimental evaluation is clearly justified when more people want a program or a service than can be helped in existing facilities (Dennis, 1994). To increase demand, the program should be publicized as part of the evaluation effort (Cook & Shadish, 1994). Unfortunately, such additional publicity could have an effect on those not selected for the program group. The control group may not act as if they had never heard about the program. Additional threats to internal validity created when a control group knows about the program are described near the end of this chapter.

Randomized, controlled experiments became more common in the 1990s, especially in the areas of community health promotion and labor economics (Cook & Shadish, 1994). The U.S. Department of Education endorsed experimental evaluations. However, internal evaluators will probably carry out experiments primarily when the situation is especially propitious, as was the case with the team versus primary nursing example above.

SUCCESSFULLY IMPLEMENTING AND INTERPRETING AN EXPERIMENTAL DESIGN

So far this chapter has emphasized that experiments have some clear advantages over other approaches to outcome evaluation and can be implemented in a variety of situations. If experiments are begun without sufficient planning, however, experiments have a way of breaking down. And, sometimes the analyses of information obtained from an experimental evaluation have been misinterpreted. To complete an experiment successfully, evaluators guard the evaluation's integrity in a number of ways (Dennis, 1994; Rezmovic, Cook, & Dobson, 1981).

Take Precautions before Data Collection

Work Closely with Stakeholders

Evaluators who work closely with stakeholders are more likely to be able to preserve an experimental design than those who assume that an agreement with an agency director is all that is needed to guarantee that evaluation procedures will be followed (Dennis, 1994). As mentioned in earlier chapters, keeping staff informed, seeking staff responses to proposed procedures, providing feedback, and being sensitive to the work of the agency will reward evaluators with greater cooperation. Those activities apply to all evaluation plans, but they apply particularly when experiments are carried out because experiments can be validly interpreted only when the groups remain intact.

Minimize Attrition

If attrition is not minimized, the equivalence of the groups that had been achieved through random assignment will be lost. Of course, attrition would not be the problem it is if it were easy to address. Still, one way to reduce attrition is to postpone random assignment until the pretest data have been collected and some attrition has already occurred. The advantage of this strategy is that only relatively committed and responsible participants become part of the experiment because they can be expected to provide the information needed for the outcome variables. This approach to reducing attrition will produce an internally valid evaluation; however, the findings can no longer be applied to the entire population the program is designed to meet. In other words, such an evaluation has limited external validity; one cannot know the degree to which its findings can be generalized to an entire population. Some early evaluations of welfare reform used this strategy, but after the information requirements were explained to participants, only a small proportion of eligible people agreed to participate. Thus, a favorable finding might only apply to particularly responsible people, not all who would be eligible for the program if widely implemented. On the other hand, if a treatment is controversial, a negative evaluation using only the participants who are the most likely to complete a program may make an important contribution. Fuller et al. (1986) studied the effectiveness of disulfiram as a treatment for alcoholism using only those alcoholics thought most likely to benefit—612 of the 6,626 who requested treatment. When the evaluation revealed that disulfiram provided little advantage even to this select experimental group, the value of the new treatment was severely challenged.

Other approaches to minimizing attrition include offering incentives for completing the program and outcome measures on schedule, obtaining the names and addresses of relatives of participants who would know where to locate participants who move before providing data, devoting time and effort to seek out participants, and being persistent in following up missing participants (Cook & Shadish, 1994; Shadish, 2002). Including some variables for which there will be no attrition, such as information from agency records (Dennis, 1994), is a good precaution.

Plan a Statistically Powerful Experiment

Stakeholders, government oversight offices, and foundations expect statistical analyses of evaluations. When planning an experiment, one needs to choose a size of sample that permits appropriate statistical analyses. The choice of sample size need not be a guessing game. First, as mentioned earlier, it is necessary to learn how much improvement would justify the expense of the program. This is a question for program staff and managers, but it is not easy to answer. However, once the outcome measures have been selected, evaluators can ask stakeholders, "What is the smallest difference in scores between participants and nonparticipants that would be enough for you to believe the program is valuable?" To help in this discussion, evaluators might present the findings of other evaluations of similar programs and use the findings of the review prepared by Lipsey and Wilson (1993) to provide a context. After stakeholders agree, one can use the standard deviation of the measure to calculate the minimum effect size they are saying would justify the program. Suppose that it is agreed that a reduction of 10 points on a widely used depression checklist would be acceptable. If the checklist has a standard deviation of 22 among moderately depressed people, the effect size that is desired is 10/22, or 0.46. This value is about equal to the effect size from the typical program studied by Lipsey and

Wilson.

With the effect size in hand, one can use power tables (see Cohen, 1987; Lipsey, 1990) to choose the sample size needed to have a good chance of finding a statistically significant difference if the outcomes match or exceed the minimum effect size said to justify the program. See the eResources for links to online assistance with power analysis, but here is an example of deciding what a good sample size might be. If a design includes only postprogram observations from a program group and a nonprogram group, the analysis will be an independent groups *t*-test. To have a 50% chance of detecting an effect size of 0.46, one needs approximately 27 people in each group using a one-tailed test of the null hypothesis (using *p* of 0.05 or less as the criterion of statistical significance). In other words, if 54 people are available for an experimental evaluation and if the participants change almost a half of a standard deviation on the outcome criterion, sampling error will lead to only half of similar evaluations being statistically significant. Perhaps the stakeholders would want a better chance. The sample size needed for an 80% chance to detect an effect size of 0.46 is 60 in a group, or 120 people altogether. The biggest challenge in this process is to know what effect size makes the program worthwhile. Staff members frequently expect bigger differences than are indeed found in evaluations of similar programs. And it continues to be true that most evaluations end up with lower statistical power than was planned due to factors beyond the full control of the evaluators.

Keep Track of Randomization While an Experiment Is in Progress

Once an experiment has begun, evaluators remain in close contact with the agency sponsoring the program to check that the randomization procedure is being carried out and that the program is implemented as planned. Furthermore, there are additional threats to internal validity that can develop after an experiment has begun.

Diffusion or imitation of the program can occur when people in control groups learn about the program and, in effect, administer it to themselves. In an evaluation of Students Against Drunk Driving (SADD), students at one of the schools included in the non-SADD control group spontaneously formed a SADD chapter during the course of the two-year evaluation (Klitzner, Gruenewald, Bamberger, & Rossiter, 1994). The analysis was thus a comparison of a school with SADD with another school with SADD—obviously, not a good test.

Attempts to compensate the control group occur when staff members seek to give special services to the people in the control group in order to make up for their not receiving the program. The efforts by nurses to give premature infants pure oxygen even though they were assigned to the regular air group were mentioned above. In other evaluations, staff members have provided additional attention to control groups. Attempts to compensate the controls have the effect of reducing the differences between treatment and control groups—thus making the program look less effective; in an extreme situation, compensation of controls could discredit a program that would have proved worthwhile if the experimental evaluation had not been tampered with.

Some control groups have felt a *rivalry with experimental groups*. *Resentful demoralization* can occur when controls come to believe that the members of the experimental group have an advantage over them and, as a result, the people in the control group reduce the effort they might have expended had they not become aware of the program. This can easily occur when a job-training program is offered to some unemployed people. People without the program could feel that in a tight job market those with the program experience will get the available jobs; consequently, controls may believe that it is not worth the effort to look for work. When a control group feels demoralized, its members may expend little effort to resolve their own problems; this reaction would contribute to making the program seem more effective than it really is.

Another issue of concern to evaluators trying to maintain an experiment is the possibility of an unplanned event occurring exclusively to either the control or treatment group that affects outcomes. Cook and Campbell (1979) call these events *local history*. For example, in an evaluation of group therapy particularly good interpersonal chemistry may make a group especially successful regardless of the approach to group therapy being evaluated. By contrast, some unique interpersonal conflicts among group members might make the same therapy seem ineffective or even destructive. The way to minimize the effect of local history is to provide the program to and assess its effectiveness in many small groups (or even to individuals if appropriate), not just to one or two large groups. If small groups are isolated from each other, the possibility of imitation of the program, compensation for the controls, rivalry with the program group, and resentful demoralization

affecting the evaluation may well be reduced as well.

A review of these threats suggests a general concern of evaluators who are focused on outcomes: quantified outcome data do not reveal very much about the day-to-day functioning of programs. Thus, outcome evaluations are most effective when they are supplemented by program monitoring to estimate the degree to which programs have been implemented as planned (see Chapter 7). Furthermore, evaluators often include observations of programs in order to develop an understanding of the quality of the interactions between staff and participants, the degree to which programs fit the needs of the participants, and how programs fit into the overall organizations sponsoring them or the community in which programs are offered. Such issues are addressed by using qualitative methods in conjunction with quantitative outcome evaluation methods.

Analyze the Data Reflectively

There is a great deal of misunderstanding about statistical hypothesis testing (see Nickerson, 2000). Many undergraduate and some graduate students pass their statistics courses believing that when a *t*-test (or other test of statistical significance) is not statistically significant, they have evidence that the groups are the same or at least any differences between them are so small as to be trivial. As shown above, if sample sizes are small there is a very good chance that an effective program will not lead to a statistically significant result. When evaluation designs with little statistical power are planned and analyzed, it is wrong to conclude that there has been program failure. Regardless of whether the null can be rejected or not, the best statistical information to report is the 90% confidence interval for the outcome variable. The 90% rather than the 95% confidence interval is suggested because in program evaluation we usually hypothesize that the participants will be better off than the nonprogram group. Such a hypothesis should be tested using a directional hypothesis test. Let's imagine that the depression checklist was used with 24 people in each group and the evaluation found a difference between the means of the two groups of 9.4 points favoring the program group. If the pooled standard deviation had been 22, the 90% confidence interval would be

$$(M_t - M_c) \pm t \times (\text{standard error of a difference between two means})$$

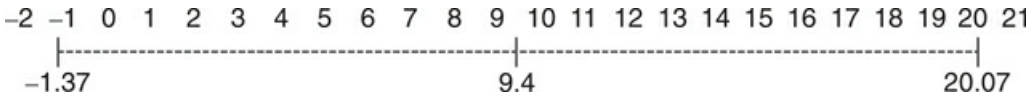
$$= (M_t - M_c) \pm t \times (s_{\text{pooled}}) \times \sqrt{\left(\frac{1}{n_t} + \frac{1}{n_c}\right)}$$

$$= 9.4 \pm 1.68 \times 22 \sqrt{\frac{2}{24}}$$

$$= 9.4 \pm 10.67$$

$$= -1.37 \text{ to } 20.07$$

Perhaps it would help to rewrite the confidence interval graphically:



What does this confidence interval tell us? It tells us that the difference is not statistically significant because the low end of the confidence interval extends below 0.0. It does *not* tell us that the program failed to make an improvement; there is no more reason to believe that the true difference is less than 0.0 than to believe that the true difference is more than 18.7 (i.e., 20.07 – 1.37) points. What can an evaluator say? When faced with a finding that is not statistically significant such as this one, evaluators should say this:

1 A difference favoring the program group of 9.4 points was found; this difference is less than the 10.0

points that was deemed the minimum to justify the expense of the program.
2 The *t* was not statistically significant; this will disappoint our funding agency.
3 The confidence interval indicates that with our small samples, the difference between the groups could be trivial or it could far exceed the minimum needed to justify the service.
4 An effect size of 0.43 (i.e., $g = 9.4/22$) was found; this effect size is marginally smaller than the minimum believed necessary, 10/22 = 0.45.

In summary, we must remain cautious. The evaluation showed there is a chance that this program is sufficiently effective; however, the samples were small and the difference found was not as large as needed to provide a good chance to detect even the minimum difference needed. We should continue to monitor this program.

These points may sound like weasel words, but they are accurate statistically. And they do not label a program as ineffective when there is a greater than 50% chance of a Type II statistical error (see Abelson, 1995; Posavac, 1998). Suppose an evaluator had been told, "We wanted a mean of 10 points improvement, but we also don't want to drop a program that is effective unless we are really sure that it is not." The analysis above would support continuing the program, but the evaluation should also be continued.

As noted earlier, although the vast majority of evaluations have not used experimental designs, especially random assignment of participants, there is growing demand for such elements, especially to address issues of what caused any change observed in programs. Case Study 11 describes an evaluation with a randomized clinical trial that provided a foundation for assessing cause. Like several other case studies in this book, this evaluation is ongoing, so developments, including additional findings, will be posted regularly in the Chapter 11 folder of the eResources.

CASE STUDY 11

Restorative Justice with Juvenile Offenders

Dr. Ida Dickie, the Director of the Forensic Emphasis Area in the doctoral Clinical Psychology program at Spalding University in Louisville, KY, has been working for years to build restorative justice practices in Louisville. As the name indicates, restorative justice involves working to restore conditions and relationships following offenses rather than only seeking to exact punishment, because it views crime primarily as a violation of a relationship rather than primarily as a violation of state law. Among other things, conventional criminal justice practices have little or no provision for victims' needs and desires. Instead of focusing only on how offenders "repay" the state with fines and incarceration, restorative justice typically provides victims with a better sense of resolution, including some ways to repair the harm done both in terms of money or goods as well as relational points such as an apology.

The project being evaluated was a restorative justice process for juvenile offenders called Family Group Conferencing, offered as an alternative to court proceedings. The process involves a facilitator meeting with family members, supporters, and the parties directly involved on both sides of an offense—those harmed and those causing the harm. Separate meetings help prepare for a full meeting of all parties where details of the events are clarified, especially the nature and extent of the harm, as members work together on an agreement about how to repair the harm. Although family group conferences do not have to be alternatives to court, that was the case with this project partly as an incentive for those eligible to choose to participate.

Experimental designs are quite rare in evaluations partly because they typically involve vastly greater time commitment, such as lead time to line up the logistics for random assignment of offenders and to bring multiple stakeholders into sufficient agreement, as well as greater time managing a more delicate process. With the high stakes of considering the legal ramifications of suggested changes to juvenile courts, however, the greater confidence provided by an experimental design seemed necessary to convince many of the stakeholders that any valuable outcomes observed could clearly be attributed to the project.

Meeting Needs

Although one need that occurs to many observers is reduced recidivism for offenders, the Louisville Restorative Justice Pilot Project took an innovative approach, recognizing that victim satisfaction and offender satisfaction with the justice system were also important needs to address. Further, the need for restoration and healing in multiple ways following offenses meant that attending to forgiveness and the details of the process for all those involved is essential.

Implementation

At the time of writing, 91 cases have been initiated. The average age of the offender is 15, and 70% are male. Regarding initial risk to recidivate, 40% were low risk, 45% were moderate risk, and 15% were high risk.

Stakeholders

As noted above, one of the important challenges for both the project and the evaluation was the range of stakeholders with strongly held positions about many of the key elements, including judges, prosecutors, defense attorneys, police, and the Department of Juvenile Justice. Multiple groups within the established criminal justice system are highly invested in particular approaches and practices. As the plans needed to be reasonably acceptable to all of them, it took substantial time for those involved in the leadership of the program to build a framework where those who normally opposed each other were able to find common ground and shared goals.

Side Effects

Part of the strategy for supporting this intensive process before the evidence of success could bring in better funding was to have doctoral students with an emphasis in forensic clinical psychology provide substantial staffing as part of their training. At one point, the stu-dents lost the data for five cases. Some students, however, saw this work as a moderately important priority rather than one of their highest priorities, the perspective of most employed professionals, due to many factors, such as time between cases, other demands, challenges with data management, and the time commitment of writing. In addition, changes in relevant laws made it more difficult for the project to have direct access to recidivism data. The resulting slower progress and fewer participants lead to a backlash from some supporters—an important negative side effect.

On a positive note, the evaluation helped to expose many stakeholders repeatedly to the details of the process, so many who were only partly aware of both how involved the intervention is and how powerful the effects could be when it was particularly successful became much more enthusiastic supporters.

Improvement Focus

One aspect of the design was a multi-wave check-in with all participants, one goal of which was to teach the process of restorative justice, a novel approach for many involved in the project. Along those lines, the multiple check-ins provided more detailed information about the extended process, both positive elements and challenges, which provided substantially greater insight for improvements. Unfortunately, the commitment to improvement on the part of the evaluator was greater than the commitment of some of the participants, so it was much harder to obtain full data from the various check-ins. Still, one of the initial findings is that more sustained interactions among the parties (a greater dose) appear to be helpful.

Outcomes

As noted, this project is still in process, so the evaluation currently addresses implementation and outcomes in process rather than the full impact. Still, before the conference hearing, 85% of the offenders said it was important to tell the victim what happened, 76% said it was important to do something to repair the harm, and 76% said it was not likely they would return to the justice system, and there were similar proportions afterwards. Before the conference hearing, 80% of the persons who were harmed said it was important to meet with the person who harmed them, and after the conference hearing, 94% said it was helpful to meet with the person who harmed them. Only 34% stated that understanding why the offense happened was more important than the offense itself.

Looking ahead to the measures of longer-term outcomes, although important progress has been made over time refining this restorative justice process, correctly quantifying the crucial elements remains an important challenge. For example, although reducing recidivism is a central goal for many stakeholders, the substantial time lag before it is clear whether participants will reoffend complicates the clearest assessment of ultimate outcomes. Still, using state-of-the-art risk assessment instruments as noted above as well as measures of forgiveness and related components will provide a more comprehensive picture of the outcomes than only relying on recidivism. The baseline figures were noted in the Implementation section. Changes over time are not yet available, but will be posted in the eResources on at least a yearly basis.

Nuances

As noted in the outcomes, there is growing insight into the active ingredients of the restorative justice process. Careful observation of the conferences was intended to assess the degree to which certain elements were important mechanisms of the good outcomes. Two of the possibilities are apologies and the level of emotionality, which is an indicator of the genuineness of the interactions. Results of these analyses will also be posted in the eResources.

Summary

This project and the evaluation illustrate that restorative justice was not designed primarily to reduce recidivism but to restore relationships, as crime is viewed as a violation of relationships. This restorative justice view is in sharp contrast to traditional justice approaches that hold reduced recidivism and public safety as the primary goals. Therefore, it is clear, more work is needed to ensure that addressing the risk factors that place juveniles at risk of future involvement with the law becomes a piece of after-care programs to complement the restorative justice process.

Sources: I. Dickie (personal communication, December 13, 2017); Duncan, S. H., & Dickie, I., *Prev. Res.*, 20, 11–14, 2013.

EVALUATOR PROFILE 11

Ida Dickie, PhD—Committed to Restorative Justice

Dr. Ida Dickie is a Justice Reformer—working to promote a radically different paradigm than the punishment-focused approach most common in the criminal justice system. Her framework is based on the principles of restorative justice and on her training as a forensic psychologist and scientist. In addition, the foundation of her approach is the value of treating others with respect in a system that too often treats people with "a disdain for the human condition," which she indicates actually means "zero respect." She attributes some of this perspective to the influence of her mother, who insisted that Ida think through how things seemed from others' point of view. In addition, she recounts how, during her school years, after she talked with one of the unpopular students, other students told her she was "not allowed" to interact with such people. In reaction, Ida deliberately chose to spend time with the "outcasts."

Although her value of restoration is clearly important, her commitment to scientific principles is equally important—as she teaches her students, one must look for clear evidence, not simply assuming that values translate into good results. Program evaluation thus fits in with her commitment to rigorous research and to allowing evidence to be the basis of decisions. She notes that, although she believes in sharing opportunities with students, those students "need to be committed to research over the long haul and to the integrity of data collection. That requires substantially greater supervision," a much more demanding form of instruction. That commitment also led to the drawn out and demanding work of building a true experiment to evaluate the Family Group Conferencing project.

These elements have specific implications. Her commitment to restorative justice means that she values healing for the various parties involved in criminal offenses, not simply evidence that the offender "paid a debt to society," a traditional definition of what happens through incarceration. For example, those who are harmed by offenses have especially important needs that are not addressed by prison sentences for the offenders. Further, she believes in broader settings, "treating each other with respect requires hearing the other's side." She acknowledges the difficulty in maintaining such an approach when dealing with some offenders. In her own experience, working with perpetrators of sexual violence was extremely challenging. But she notes that "I never could have developed this tolerance and understanding for the negative sides of the human condition" otherwise. Doing so has given her "a sense of respect" for herself, as she is committed to "what I see as my purpose, my life's mission, to bring this level of respect to when we really, really hurt each other."

Source: I. Dickie (personal communication, December 13, 2017).

SUMMARY AND PREVIEW

Whenever program evaluations are designed to learn whether or not a human service program caused positive changes in the people served, an experimental evaluation design should be considered. It should not, however, be forced on a program or facility that does not support the requirements of experimental research. When completed, program evaluations based on experimental research designs are easier to interpret and to analyze than are the less well-controlled evaluations described in Chapter 10. However, even the most carefully designed evaluations are not interpretable if they are not carried out as planned. Evaluators recognize that many human service professionals do not fully understand the advantages of the experimental approach and could in all innocence reduce the validity of an experimental evaluation by providing extra care for the control group before the outcomes are measured. As with all forms of evaluation, someone needs to monitor the way people are assigned to groups, how the data are collected, and how the program is implemented.

Few program outcomes are worthwhile regardless of costs, and many socially desirable objectives can be approached in more than one way. Consequently, costs are frequently considered when the worth of program outcomes is considered; to do that cost–benefit and cost-effectiveness procedures, to be described in Chapter 12, are used.

STUDY QUESTIONS

1 List the advantages and disadvantages of experimental and quasi-experimental designs in program evaluation. You might use parallel columns so that when you are finished, you have a concise way to compare these two approaches. This effort could help you on a test.

2 Using language that can be understood by someone without training in statistics and research methods, explain the advantages of experimentation, including random assignment to groups. Imagine you are addressing elementary school teachers, nurses, or police officers. The adequacy of your answer depends on both technical accuracy and likely audience understanding.

3 Suppose that after an experiment has been described (as in question 2), Brett asks, "If the new program is worth the time and effort to study it, someone must think it is worthwhile. And if the staff believes that the new program is good, how can you justify not letting everyone benefit from the program? Having a control group is unjust. And it takes too much time." How would you respond to Brett?

4 Austin decided to carry out a program evaluation of a drug rehabilitation program for women on probation using an experimental design. He asked probation officers to randomly assign women from their caseload to the program or to a control group that was to receive standard attention from probation officers. The program lasted 15 weeks with three meetings per week. After the program was over, the

participants were referred to a less intense follow-up program. Austin prepared rating forms for the probation officers to use at the end of the 15-week program to rate the success of the women for whom they were responsible. They were to use the forms to rate the degree to which the program participants and the control women were following a drug-free life, developing a stable lifestyle, and avoiding further legal problems. Austin planned to use these ratings as his outcome variables to compare the condition of those in the program with those in the control group. Austin followed some procedures that were almost guaranteed to make his findings uninterpretable. List problems with his procedure and what he might have done differently to produce a more credible evaluation.

5 Caleigh was a new evaluator in a clinic setting and not yet clear on her role. She learned that several short courses on pregnancy and baby care were offered to expectant mothers. Only 40% of eligible mothers attended. Because these programs had never been evaluated, Caleigh proposed an evaluation to the Clinic Medical Director. Caleigh suggested that those women indicating interest in the courses be randomly assigned to the courses or some alternative such as a workshop on family budgeting. She proposed using a checklist of problems that the new mothers experienced after their babies were born as an outcome assessment. If the courses are valuable, she said that those who completed the courses should experience fewer problems.

6 There are many problems with this proposal. First, list the ethical and design problems. Second, suggest some aspect of this program that Caleigh might profitably evaluate. What might be added to the program that would permit an ethically sound experiment? (Hint: how might expectant mothers learn about the program?)

ADDITIONAL RESOURCE

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Boston, MA: Houghton Mifflin.

This is the same resource recommended for the previous two chapters, because it is the classic in the field with no current replacement. Chapters 8, 9, and 10 are devoted to randomized experiments. Chapters 11, 12, and 13 concern the central issue for evaluators who find themselves called on to show whether a program shown to be effective can be applied in settings other than the one in which the evaluation was carried out.

12 Analyses of Costs and Outcomes

Programs have not been adequately evaluated if costs have not been considered. Even an evaluation controlling for all the threats to internal validity and finding statistically and substantively significant improvements in program participants can be a negative evaluation if the costs necessary to achieve those improvements exceed what society is willing to pay or exceed the costs of alternative approaches that can resolve the same problems. Recommendations for program development that are not accompanied by estimates of the costs of such development will often not be taken seriously. Being able to relate program outcomes to costs is becoming increasingly important to evaluators who want their findings to be used in agency decisions. Concerns over the efficiency of human services have become very widespread (Fals-Stewart, Klostermann, & Yates, 2005; Ross, Barkaoui, & Scott, 2007). Unfortunately, service providers and evaluators seldom have training in cost accounting, and those who work for nonprofit organizations are remarkably naive about the costs of providing services (see King, 1994; Yates, 1994). If you eavesdrop on college faculty members discussing higher education budgets, you will discover that many faculty members know very little about how a college funds its activities.

Cost analyses, however, are not novel among business managers. The local hardware store manager counts receipts every evening and summarizes costs (of course, computers handle these activities). By converting transactions into dollars, managers can compare costs (salaries, inventory, and buildings) with cash receipts. A comparison of these figures tells managers whether they are making or losing money. Of course, bookkeeping is not that simple; however, the principle is straightforward and the unit of analysis, dollars, is widely accepted.

In human service settings the principles are less clear, and the benefits are usually hard to convert into dollars. Further, there typically are multiple elements of programs whose value cannot be easily assigned a dollar figure, so comparisons are complicated (King, 2017). Nevertheless, cost projections and program objectives must be made and used to choose among possible programs before anything is implemented. Program planners seek to propose services that will merit the investment of funds. After a program has been offered, evaluators assess the quality of the program's implementation and impact. To complete the evaluation, the outcomes can be related to the funds expended using the methods described in this chapter.

PROGRAM COSTS

Types of Costs

Placing costs into categories helps evaluators recognize what is needed to support a program.

Fixed versus Variable Costs

There are costs that must be borne simply to open an agency's doors to the first client. If car mechanics are to be trained, the cost of facilities and tools must be covered before even one person enrolls. Such costs will be the same whether 2 or 22 people are trained. On the other hand, there are variable costs that depend on the number of program participants. The costs of supplies and maintenance will increase with additional students. Universities have high fixed costs due to large physical plants and tenured faculties. Enrolling a few more undergraduates adds very little to variable costs; even a 5% or 10% increase in enrollment does not increase the university's expenses very much due to the large fixed costs. On the other hand, falling enrollments and large fixed costs are very hard on university budgets.

Incremental versus Sunk Costs

Incremental costs are those that must be covered day by day in order for a program to provide a service—staff salaries or repairs, for example. Sunk costs are those that have been expended already. The difference between the original cost of a car and its current value is a sunk cost. The reason to make this distinction lies in the difficulty that people have in understanding that sunk costs should not determine future behavior; incremental costs should. Two years after the U.S. Postal Service installed a billion-dollar package sorting system, the GAO recommended that the system be scrapped (Postal Service, 1990). The costs of maintaining it—incremental costs—were more expensive than the alternate ways of sorting packages. The original cost of the equipment was simply irrelevant to the decision about the future because incremental costs were going to exceed the benefits that would be obtained by continuing to use the system. People find it very difficult to "cut their losses," but in some circumstances the losses will simply continue in the future. Choosing to abandon a failed policy or expensive, but inadequate, equipment is often a difficult decision to make, even when not doing so would mean that losses will increase as time goes by. But such a decision is the one supported by the evidence.

Recurring versus Nonrecurring Costs

As the term suggests, recurring costs are those that are due at regular intervals, such as salaries, rents, and utilities. Equipment or buildings that are expected to last a number of years before being replaced are nonrecurring costs.

Hidden versus Obvious Costs

This distinction is presented just to highlight the point that some costs are easily recognized, but others easy to overlook, especially when service is expanded. For example, a plan for a building addition always includes the cost of the new structure. But there have been times when planners forgot to consider furniture or custodial expenses. When a program expands its staff, not only must additional salaries be added to the budget, but fringe benefits and liability insurance costs must be taken into account as well.

Direct versus Indirect Costs

Another way to divide costs is to consider those that are associated with providing services and those that support the agency that makes it possible to provide services. The salary and fringe benefits of a social worker make up the direct cost of providing service to a specific client. However, in order to provide that service, the agency must be managed, receptionists must make appointments, telephones must be available, and buildings must be cleaned. All of these expenses (and more) are incurred if the agency is to remain operational. Often indirect costs are called *overhead*. When estimating the costs of providing a program, planners must consider both direct and indirect costs.

Opportunity Costs

A mistake often made when considering a new program is to ignore the necessity of making choices among alternative uses of funds and effort. By adopting program A, it is impossible to adopt a different one. What cannot be done is called an *opportunity cost*—a lost opportunity, if you will. The most costly factor in attending college is not the tuition—the most costly factor is the income students cannot earn while attending college. In time, most college graduates more than make up for the cost of tuition and the foregone wages. However, in other settings the selection of one program makes it impossible for an agency to serve another purpose. There are usually many promising activities that cannot be supported within a budget even though all of them are needed and contribute to meeting the mission of the agency. Whenever money is spent to relieve one need, it cannot be used to relieve some other need, an opportunity cost. In debates about plans for agencies, many proponents of particular emphases tend to ignore the activities that cannot be supported if the preferred

emphases are supported. Any choice you make involves an opportunity cost, but clearly opportunity costs are not the same sort of costs as paying for the program chosen. It's like getting a rocky road ice cream cone and thereby foregoing the pleasure of butter pecan; there is a cost, but not a cost of money.

An Example Budget

Figure 12.1 gives the budget of a small Special Opportunity Program for underprepared college students. The proposal writers assumed that a small furnished office could be found in an existing building. The writers did not include maintenance, custodial services, or insurance costs because these would be the same regardless of the program the university decided to support and house in the space. University administrators would have those estimates if needed. The budget covers the particular costs that would be associated with this program. University administrators would decide whether this program helps fulfill the mission of the university better than some other activity that might be housed in that space.

Salaries (Employees)	
Director (half-time faculty)	$ 45,000
Secretary—Office Manager (half-time)	$ 24,000
Fringe benefits (17.5% of salaries)	$ 12,075
Stipends—Student Assistants	
3 Advanced Tutors	$ 51,000
3 Basic Tutors (Work-study undergraduates)	$ 5,400
(University contributions 50%; state contributions 50%)	
Total Compensation	**$127,475**
Office Support	
Supplies	$ 4,500
Computer and laser printer	$ 3,000
File cabinets	$ 1,100
Copying and printing	$ 2,000
Phone	$ 1,800
Total Office Support	**$ 12,400**
Program Total	**$149,875**

Figure 12.1 Proposed first-year budget for a Special Opportunity Program to serve 35 college students.

The Necessity of Examining Costs

Public and private agencies must consider costs. Since funds are limited, choices are always being made. A rural Arkansas county instituted services for elderly residents to improve the quality of their lives and to

provide services that would postpone admissions to nursing homes (Schneider, Chapman, & Voth, 1985). The evaluators discovered that although most elderly residents thought that having the services available was a good idea, only 12% came to the center once or more often per week. Depending on the accounting assumptions adopted, the cost to run the center *per user* per year was 58%–71% of the cost of educating one schoolchild for nine months. While those who used the services benefitted, this center seemed to cost more than the value of the services provided.

Although the Arkansas senior center cost analysis was part of an evaluation of an operational program, planners also try to project costs of policies and programs before implementation. Unfortunately, many guesses are grossly in error. Sometimes the program does not attract as many users as estimated, and sometimes far more people seek a service. In 1973, Congress passed a bill to pay for kidney dialysis through Medicare for people suffering from kidney failure and unable to obtain a kidney transplant. When the bill was approved, the costs were estimated to be in the low millions. The cost to Medicare for kidney dialysis was $9.2 billion in 2007 ("Medicare proposed payment," 2009). This is not to suggest that this is a bad way to spend federal funds because dialysis extends the lives of people. The point is simply this—it is hard to provide an accurate estimate of future costs for a program. These high costs were not expected and make it impossible to deal with other pressing health concerns. Not meeting other needs is an opportunity cost that raises important questions of fairness and justice.

COMPARING OUTCOMES TO COSTS

Determining costs is one part of the job—although seldom easy, it is the easier part. The more complex task is estimating the benefits of a program. Benefits occur when objectives are achieved. Surgeons prevent premature deaths when they remove diseased appendixes. Because their patients lead productive lives afterward, there are considerable benefits to this and many surgical procedures. Increased earnings are a benefit of college completion (Gross & Scott, 1990). It is likely that benefits accrue to the friends and family members of participants in drug use prevention programs (Caulkins, Rydell, Everingham, Chiesa, & Bushway, 1999). French (2000) found favorable returns for alcoholism treatment services.

Another way to compare costs and benefits is by first calculating how much money one would have in the bank if the cost of the program were saved rather than spent on the program. Imagine the amount available if a person saved the money that could have been earned instead of attending college (less living expenses) plus the cost of tuition, books, and fees. Assume that this money is invested. Knowing the wages of different occupations, an economist can contrast the average lifetime wages of college graduates with those of workers who completed high school only. A person who began working immediately after high school would have a sizable account by the time a college graduate began a full-time job; however, the college graduate would be earning more per year. Would the employed high school graduate remain ahead? It is possible (using techniques not covered in this book) to calculate the rate of interest that the high school graduate would need to get on the amount saved (by not attending college) so that the investment would equal the benefit of attending college.

Using these methods, it turns out that over a 30-year career a high school graduate would need to find a source giving 34% per year on the investment in order for it to make financial sense to go to work right after high school in the 1970s (Sussna & Heinemann, 1972). Since few investments can match that rate of growth, such analyses mean that financial benefits of college graduation exceed its costs. Similar analyses show different percentages, but lead to the same conclusion—college is worth more than the cost (Sanchez & Laanan, 1997; for a similar European-based study, see Hartog, 1999). However, it might not be worth the *extra* cost to attend an elite university rather than a less expensive but still good one (Kahn, 2000).

Note that even in this simple example of calculating the economic benefit of completing college, there are many assumptions and estimates that must be made. For example, job security is often higher for college graduates than for those beginning work after high school. Furthermore, the economist must assume that the types of jobs available would remain about the same as in the recent past. When discussing any complicated analysis or policy, it is said, "The devil is in the details." It is critical to ask about assumptions whenever benefits are related to outcomes.

An Illustration of Cost–Benefit Analysis

A cost–benefit analysis compares the cost of a program with the benefits projected to result from that program. A promising method of improving the reading and math skills of elementary school children is called "rapid assessment" (Yeh, 2006). The general idea is that learning is facilitated when the learner has frequent feedback on learning. The frequent, computer-administered tests inform the learners and teachers of progress, lead to more appropriate teaching, and encourage a sense of accomplishment. In a variety of evaluations, the effect size of a year of classes using rapid assessment is approximately 0.30 more than standard school procedures.

A benefit–cost analysis of the implementation of rapid assessment throughout the United States can be used to illustrate the essence of this analysis. The technical details cannot be covered here, but the critical steps can be examined (Yeh, 2009). There are two aspects to the costs of the program: the added cost of the computers and programs for the tests and teacher training. These costs were summarized on a per-student basis. The higher levels of accomplishment for students learning reading and math using this approach mean that more students will attend college than would have otherwise; thus, they will pay more taxes in the future than they would have had they not benefited from the program.

The benefits of this method of teaching lead to greater incomes, lower levels of crime, lower welfare costs, and higher levels of tax revenues for states and the federal government. These are not estimates that can be made on the back of an envelope; however, using research showing lower rates of crime and higher levels of income among those with greater educational levels, such estimates can be made. The benefit-to-cost ratio was carried out for all 50 states. Here are the estimates for Tennessee:

Cost of program for 74,196 students was $441 million.

Benefits to the state government over the lifetimes of the students:	Increased income tax	$109 million
	Increased sales tax	$3,288 million
	Welfare savings	$2 million
	Prison savings	$17 million
	Total	$3,416 million

Benefit divided by cost: $3,416 million/$441 million = 7.7.

In other words, for every dollar that the state devoted to implementing rapid assessment in the schools, the state would receive a benefit of nearly eight dollars. Of course, the students themselves would enjoy a higher standard of living, which Yeh called "social benefits." By the way, note that Yeh has assumed that greater number of college-educated people would find jobs similar to those found by college-educated people in the past. At some point, society would not have enough additional jobs. Yeh also assumed that the effect of rapid assessment could be achieved for five years (for a combined effect size of 1.50) and that the gains could be maintained through high school. Such assumptions may or may not prove to be accurate.

Rajkumar and French (1997) examined efforts to estimate the intangible costs (e.g., increased pain, suffering, fear) of the crimes committed by people using drugs. Because drug use prevention programs reduce criminal activity, such reductions should be included in the outcomes of drug use interventions. Doing so would show that drug use prevention programs produce benefits that clearly exceed the programs' costs. The Army Corps of Engineers has routinely used the estimated recreational value of water projects to help justify construction costs. To calculate the benefit-to-cost ratio of a safety program that reduces on-the-job injuries, it is necessary to estimate the dollar value of the extended working lives of employees. If a job-training program permits some participants to leave welfare, benefits may be estimated from wages earned. But what is the dollar value of better mental health? Of reduced anxiety among cancer patients? Of increased art and music appreciation? Clearly these outcomes are valued, but how much they are worth is subject to disagreement. Cost-effectiveness analysis was developed to use in situations in which the dollar values of

benefits are very difficult to quantify and, yet, a choice among alternatives must be made. As you can see, depending on the specific estimates of value, different conclusions may be reached. It is important to present the results of these analyses with appropriate caution, especially when the basis for the estimates is less than certain.

The Essence of Cost-Effectiveness Analysis

There are outcomes for every effective program. When expressed as in numbers, it is possible to ask what it costs to achieve the outcomes of the program. Many schools screen children for health problems. The screening effort costs the school district some resources. It would be helpful to relate program costs to the degree objectives are achieved. For example, children in poor areas might be screened for emotional problems because in many families such problems would not be detected. The objective is to find students who could benefit from emotional support and therapy. The number of students identified and successfully linked to services could be the measure of effectiveness. Dividing that number into the cost of the program would give us the cost of each successful referral. In one school district, it was found that the cost of screening was between $8.88 and $13.64 per student. Depending on the level of need in the school, those values then translated into between $106 and $417 per child referred for therapy (Kuo, Vander Stoep, McCauley, & Kernic, 2009). When expressed this way, school boards have a better way of deciding whether to begin or maintain a screening program.

The power of cost-effectiveness analysis becomes especially evident when two or more programs result in an improvement of the same behavior or reduce the frequency of the same problem; the cost of a success can be found for each program, thus identifying the most cost-effective program. Suppose an existing smoking cessation program costs $1.4 million and leads 1,765 smokers to quit for at least five years. Suppose further that an alternative program is proposed to replace the current one: it would cost $5.5 million and improve the outcomes. On the basis of past research, this proposed program is expected to lead 2,728 smokers to quit for at least five years. The outcomes are better, but the costs are higher. The current program costs $1.4 million/1,765, or $793 per success. The proposed program costs $5.5 million/2,728, or $2,016 per success; thus the current program is more cost-effective because it costs less for each success (see Herman, Avery, Schemp, & Walsh, 2009). Note that sometimes analysts express these findings by dividing effectiveness by costs, the reciprocal of the values above. It does not matter which approach is used; it is important to remember whether dollars per one good outcome is found (as above) or number of good outcomes per $1,000 is found.

Cost-effectiveness analyses can be used in different ways in planning and managing an agency (Peterson, 1986). First, if a *set amount of money* is made available to alleviate a need, an analysis can be done to learn which way of spending the funds is expected to achieve the best outcome for the funds available. For example, suppose the goal was to reduce highway fatalities. This goal could be met by spending the specified funds on more police patrols, on an educational program to increase seat belt use, or on construction to eliminate some dangerous turns on rural highways. The approach that promised the best return would be chosen. Second, the costs of achieving *a specified outcome* can be compared for various possible approaches to a problem. A school district wishing to raise average daily attendance rates from 87% to 92% could consider developing a lottery for which students get one ticket for each week of perfect attendance, initiating an automated telephone system informing parents when their child is absent, or redesigning the vocational curriculum. Third, if one is required to use *a specified type of program* to address a problem, the cost-effectiveness of different intensities of single type of program might be found. A city council might, for example, ask to what degree coronary deaths would be reduced by spending $2, $4, or $6 million on paramedics and intensive care ambulances.

When Outcomes Cannot Be Put into the Same Units

In the earlier examples, the outcomes of different programs could be placed into the same units—for example, the number of highway deaths expected to be prevented could be compared for several programs. However, it is not unusual for planners to be asked to choose among programs designed to achieve different objectives, none of which is easy to put into dollars. How is Congress supposed to compare the outcomes of funds supporting space research with outcomes of Medicare services? A college dean cannot agree to support every

proposal from faculty members, some of whom want more chemistry faculty, while others want a bigger art gallery, and yet others think that an American Studies Institute should be funded. And the coach claims that with two more basketball scholarships, the college will become very competitive in women's basketball. Although the cost of these proposals can be estimated, their likely benefits to the college cannot be converted into the same units. Cost-utility analysis was developed to aid in making such choices (Levin & McEwan, 2001). In brief, cost-utility analyses require stakeholders to examine and compare the perceived desired outcomes (the subjective utilities) with the costs of these very different proposals. The hypothetical Special Opportunity Program, whose budget was given in Figure 12.1, would be compared to other proposals in subjective terms, not in terms of quantified effectiveness or the dollar values of benefits.

SOME DETAILS OF COST ANALYSES

There are a number of detailed issues to examine when conducting an analysis of costs. Although these points are subordinate to the major conceptual ideas presented above, they can affect the impressions people receive from cost–benefit and cost-effectiveness analysis. There are no standard ways of dealing with these issues.

Units of Analysis Must Reflect the Purpose of the Program

The units of analysis used should reflect the goals of the programs compared. What is the purpose of prison sentences? If the purpose is to keep prisoners off the streets, then a successful outcome is that a person found guilty is in prison for many years. Society would then find the least expensive way to house prisoners and choose that system. If the purpose is to lead guilty persons to lead productive lives and no longer be a threat to others upon release, that purpose would imply that a successful outcome is a released prisoner who finds a job and remains out of trouble. It is easy to see that those different definitions of successful outcomes would lead to quite different cost-effectiveness analyses.

Kane and Kane (1978) urged that nursing homes for elderly people be evaluated on the basis of the health and mental status of residents. Using costs per resident per day to evaluate the efficiency of nursing homes encourages the erosion of treatment standards and results in limited privacy, superficial medical care, recreational opportunities centered on TV viewing, and artless food preparation. Hospice care for terminally ill patients focuses on an end of life that is comfortable and pain free rather than a longer life accompanied with tubes and extraordinary medical efforts such as kidney dialysis (Tamura et al., 2009).

When thinking about the unit of analysis, you should remember that the impact of a program is not necessarily linear. It would probably be harder for a person to develop new job skills leading to an income increase from $0 (i.e., unemployable) to $15,000 than to improve from $4,000 to $19,000, even though these are numerically identical improvements. Helping an unemployed person to hold a job may require teaching about a job behavior as well as teaching specific job skills and, consequently, would be harder than helping an employed person to improve skills in order to qualify for a better job.

Future Costs and Benefits Are Estimated

A sophisticated cost analysis includes benefits expected to occur in the future. Successful rehabilitation for drug addicts is expected to mean that the individual will become productive and not require additional services in future years (Rajkumar & French, 1997). Improvements in work skills, psychological adjustment, and physical health may well have benefits for the children of people served. The worth of these long-term benefits is, not surprisingly, hard to estimate. Viewing the human service recipient as a member of a family and a community system is a way to become sensitive to possible benefits as a consequence of the program participant making some behavioral change as a direct result of the program. Successful drug rehabilitation should lead to less crime (Gerstein et al., 1994) because addicts frequently support their addictions through theft. If crime is reduced, commercial activity may increase since business owners will be more inclined to venture into the locations previously at the mercy of addicts. Also, a lower crime rate would mean that the community could lower the amounts spent on detection, prosecution, and punishment of criminals.

Noneconomists face problems in comparing programs requiring funds and providing benefits at different times in the lives of programs. The funds needed and the benefits available in the future are worth less than

the same amount of money and benefits available now. This principle makes sense when you ask yourself whether you would prefer $100 now or in 12 months. Ignoring for the moment our normal preference for instant gratification, it makes economic sense to have the money now because if it were put into a savings account it would be worth more than $100 in 12 months. This observation helps to explain why state lottery prizes spread out over 20 years are not worth the value that is implied in ads. The present value of the twentieth $50,000 payment of a million-dollar prize is only $10,730 when the prize is announced because that sum invested with a return of 8% per year would be worth $50,000 after 20 years. Of course, the actual rate of return matters substantially, because if the return were only 4%, the current value would be $22,819.

Calculating the current value of future benefits is called discounting and can get tricky. The point to remember is that the cost of programs should include what could be obtained by an alternate use of resources. For example, is it worth investing $100 now in order to guarantee receiving a benefit of $200 10 years from now? The answer depends largely on the assumed rate of return for an alternative use of the $100. The amount one could obtain from this alternative use is the opportunity cost. If the rate of return for the alternative use were 3%, it would be worth investing the $100 in the program (not in the bank) and waiting for the $200 because $100 invested at 3% would be worth only $131 in 10 years; the opportunity cost is less than the benefit. On the other hand, if the rate of return were 8% or more, the program would not be worth supporting because an 8% return would yield $216 in 10 years; the opportunity cost would exceed the benefit.

In order to handle the possibility of getting different results simply because different assumptions were used, sensitivity analyses are performed by repeating the analysis with a variety of plausible assumptions (Levin & McEwan, 2001). If the program plan is promising under a range of assumptions, some of which do not favor the program, planners can have more confidence in the findings. If different assumptions result in different conclusions, then even greater care must be devoted to the planning process. If low rates of return for alternative uses of the funds are assumed, smaller future benefits from the project would seem more desirable than if higher rates of return are assumed. It is probably obvious that fluctuating inflation rates and interest rates greatly complicate these analyses.

Who Pays the Costs and Who Reaps the Benefits?

The discussion up to this point has been phrased in terms of whether one should devote resources now to reap a benefit later, implying that the one who bears the costs is also the one who reaps the benefit. However, most programs are sponsored and paid for by people who do not obtain the benefits. The costs of primary education are borne by a community, whereas only the children (and their parents) and the teachers benefit. The other residents of the community at best benefit indirectly; if a community gains a reputation for having good schools that attract and hold residents, the values of homes remain high and good employees are not likely to move away. A tax increase referendum is more likely to pass if residents believe that they benefit at least indirectly by having good schools in their community. Public transportation systems are often criticized by people who are required to support them through taxes but who do not use buses and rapid transit trains.

The perspective taken will affect the conclusions of cost–benefit analyses (Yates, 1999). If governmental funds cover program costs, programs that seem worth the cost to individuals may not seem worthwhile from a societal perspective. Another example of the contrast between the stakeholders who pay and those who benefit from a policy is seen in the rapid development of outpatient surgery practices. Hospitals came under intense pressure from insurance companies and federal and state governments to lower costs or at least to limit increases. One response to this pressure was to develop outpatient surgical centers. Patients needing minor surgery do not stay in the hospital overnight as they would have 20 or 30 years ago. The cost of providing outpatient surgery is markedly less than surgery with a hospital stay. Family members are expected to take care of the patient after surgery (a cost), but the patient does not benefit from this change in procedure. The savings gained through outpatient surgery go to Medicare, Medicaid, and insurers (Evans & Robinson, 1980). Although it was hoped that society in general would pay lower costs for medical insurance or taxes when the costs of surgery were minimized, evidence suggested that surgeons raised their fees for surgery done outside of hospitals (Millenson, 1987). This negative side effect was not an outcome that planners had in mind when outpatient surgery clinics were encouraged. Do note, however, that outpatient treatment is preferred by patients and that being an outpatient reduces the chances of exposure to infections, which is an additional

benefit to patients. These points obviously highlight the challenges in determining costs and benefits of programs.

Pointing out the contrast between those who pay and those who benefit should not be taken to mean that this discrepancy is unusual. People voluntarily enter into many contracts from which they hope to receive no financial benefits. Home, health, and car insurance policies are designed to provide money to those in need after an illness or accident. The costs are borne by many who never have the occasion to use the benefits. The year-to-year benefit is the peace of mind gained by knowing that large expenses would be covered if a misfortune occurs. Most policy holders are satisfied with such assurances and would be delighted never to need to receive a financial benefit from their insurance.

Some critics of cost–benefit and cost-effectiveness analyses argue that the techniques are flawed because they are not sensitive to long-range costs. House and Shull (1988) argue that many analyses of the effects of the use of natural resources do not reflect the cost to future generations. For example, the financial value of forests would lead an owner to clear-cut the forest, sell the logs, and leave the forestland bare (Clark, 1989). The contribution of the forest to soil conservation, carbon dioxide storage, oxygen replacement, and ecological variety would not be considered by a single-minded owner because society in general bears those ecological costs, not the owner of the forest or the buyer of the wood. These costs are borne by future generations who may suffer if environmental damage leads to a lowered quality of life. Some uses of the environment may lead to negative long-term effects for future generations. When cost analyses are made explicit, it is easier to discover good and bad effects of a program or policy than would be possible without relating costs and outcomes.

Using Cost–Benefit and Cost-Effectiveness Analyses

If the benefits of a program can be priced in dollars and a cost–benefit analysis can be done, one approach to program selection is to adopt the program with the largest benefit-to-cost ratio. But many decisions involve far more than simply supporting or not supporting a program (Nagel, 1983). Because government agencies have limited budgets, agency managers are not free to choose a program with the highest benefit-to-cost ratio if the cost exceeds the allotted budget. Instead, managers ideally should seek to obtain the most benefits for the resources available. Given the costs of the various programs and the size of the budgets, the best program to select might not be the most efficient.

If the outcomes of alternative programs are expressed in different units, stakeholders and managers can still use analyses relating costs with outcomes to select which programs to maintain or expand. The crucial technique is to list all possible courses of action and compare them to each other, taking two at a time. For example, suppose that a program to train 3,000 unskilled workers per year to fill maintenance positions could be funded for 10 years for the same costs as a road improvement project that is expected to save 5 accidental deaths and 20 injuries during the same 10 years. If both outcomes are equally likely (according to past research and evaluations), which course of action should a state legislature choose? This choice can be made without estimating the value of jobs and averted deaths and injuries; however, the choice would reflect the relative importance of the two outcomes for the majority of legislators. Next, a third program could be compared with the one preferred. Thus, even when cost–benefit and cost-effectiveness analyses cannot be done, relating costs to outcomes illuminates the debate and allows the stakeholders to see options more clearly than they could without such analyses. At a time when limitations on the financial resources of governments and other organizations are crucially important, analyses to help decision makers understand the impact of their choices seem worthwhile.

MAJOR CRITICISMS OF COST ANALYSES

There are a number of criticisms of both cost–benefit and cost-effectiveness analyses. Users of these approaches profit by being aware of these criticisms, which point out limitations of these useful techniques.

The Worth of Psychological Benefits Is Hard to Estimate

As mentioned above, improved mental health is hard enough to price, but what about clean air? Clean

streams? An unlittered park? One approach is to study what people are willing to pay to escape environmental deterioration. Another approach is to calculate what must be paid to induce workers to accept employment in dirty, dusty, noisy, or dangerous occupations (Viscusi, 1992) and examining how much more people are willing to pay for property in more desirable locations (Pearce & Markandya, 1988). These approaches are indirect and may well never prove to be totally satisfactory, but they do provide some guidance.

Placing a Value on Lives Seems Wrong

People are very uncomfortable when asked to consider the idea that dollar values can be placed on lives. Economic analyses based strictly on current or projected earning power place low values on the lives of children because children cannot earn anything until 10 or 15 years in the future. Indeed, discounting future earnings is not the way to learn what value people place on lives; people do not act as though children are to be valued on the basis of economic potential. Considerable resources are expended on all children, especially sick children. A very ill, premature baby may be hospitalized for months at a cost of tens of thousands of dollars. Similarly, economic analyses fail when applied to the elderly, who are unlikely to hold any jobs and thus would appear to have a negative economic value.

Nevertheless, society does act as though lives have financial values. These values are estimated by examining what additional wages are required to hire people for jobs in which injury or accidental deaths are higher than most jobs and what people are willing to pay to reduce a risk. Elephant handlers receive extra compensation relative to zoo workers taking care of less dangerous animals (Viscusi, 1992). By examining the higher rates of job-related death among elephant handlers, one can work backwards to learn what value has been placed on a life. Here is an easier example of working backwards. The risk of dying in an automobile accident was about 1.5 per 100,000,000 miles in 2005 (National Safety Council, 2009). This means that for cars driven 10,000 miles per year, the chance that someone will die in one is 1.5 out of 100,000,000 divided by 10,000, or 1.5 out of 10,000 in a year. Of course, you lower your risk by never driving when tired or after drinking, but let's just accept the average for the U.S. population. Suppose you could spend a sum of money and reduce that risk to one-third of the average rate, that is, lower the risk to 0.5 out of 10,000 (which is equivalent to 1 out of 20,000). What sum would you pay? Think about it for a second—$50 per year? $100 per year? $200 per year? How much? If you would be willing to pay $200 per year, you are in effect valuing a life at $4 million ($200 divided by the risk you are willing to tolerate, that is, 1/20,000, is $4 million). This illustration is artificial because we are seldom presented with such clear questions; however, we make decisions daily to buy a carbon monoxide or smoke detector or not, to replace a worn extension cord or use it anyway, or to buy a bicycle helmet or do without one and save the $40. In 1992, Viscusi reported that similar decisions by workers revealed that they value a life at between $3 million and $7 million. Medicare seems to be valuing a year of life (note, "a year of life," not "a life") at $129,000 (Kingsbury, 2008).

Examining policies in terms of the costs to reduce risk can reveal unnoticed assumptions. The National Research Council reported to Congress that requiring seat belts in school buses was estimated to cost $40 million per death averted (Derby, 1989). It does not take too much imagination to think of ways to spend money that would have a better effect on children's health than placing seat belts in school buses. Suggesting alternative uses of such money would not mean that policy makers did not value the lives of children riding buses; it does mean that there are more cost-effective means of reducing accidental deaths of children. The process of learning what the effects of society's actions are does not devalue lives. Estimating dollar amounts can appear to devalue life if the critic assumes that there is an infinite reservoir of funds to protect people from every possible risk. At times some people act that way, especially when they demand others to bear extravagant costs to reduce specific risks. However, public resources are not infinite; choices must be made among a variety of good things. It has been suggested that *less* effective medical treatment be used at times because the cost of the most effective would be unacceptably high (Kent, Fendrick, & Langa, 2004).

Cost–Benefit and Cost-Effectiveness Analyses Require Many Assumptions

Whenever future costs and benefits are estimated, assumptions and projections must be made. Since there is a great latitude in selecting assumptions and making approximations, it is not surprising that cost–benefit and cost-effectiveness analyses can be used to support predetermined opinions (Joglekar, 1984). Being open about

assumptions and giving the reasons behind approximations enables others to assess the validity of the analyses. Furthermore, adopting different assumptions and discount rates permits evaluators to offer a range of conclusions. If such a range still supports the policy, one can have increased confidence in the conclusions. Cummings, Rubin, and Oster (1989) reviewed the literature on the effect of physicians encouraging patients who smoke to quit; they concluded that even if only 1% of smokers quit after such discussions, such efforts are more cost-effective than treatments for mild hypertension and some other common preventive medical activities.

Conducting a sensitivity analysis is critical. If the program appears to be cost-effective regardless of the specific estimates made, then the conclusions take on more credibility. If a program can be shown to be cost-effective only when rather tenuous assumptions are made, then the analysis may be greeted with skepticism (Joglekar, 1984). Readers still uncomfortable with placing dollar amounts on relieving personal suffering or preventing accidental deaths should remember that financial considerations are not the only basis of program selections. But financial issues cannot be ignored.

CASE STUDY 12

Economic Evaluation of KIDS NOW

A common pattern with case studies in many fields is that they represent situations that are easier rather than harder to address. For example, the first author of this text chose the evaluation reported in Case Study 5 partly because his efforts to resolve the challenge were so quickly met with good will on the part of the other evaluators, and the results were very positive. Although other cases did not have perfect settings, many of them were generally favorable for the evaluation.

Case Study 12 presents a real-life case that shows many good aspects, but the setting of the evaluation had far more challenges and difficulties than usual. There were substantial limits on the information available, especially because, as primary and secondary preventions, the focus was on providing broad services to large numbers of people rather than on expending resources on systems and staff to document those services. The evaluators were clear with those requesting the evaluation that some of the understandable desires, such as showing clear evidence of the full impact of the intervention, simply could not be met. In addition to the challenges with tracking some of those served, when interventions are addressed to broader populations to prevent problems from arising, it is almost impossible to document which cases would have shown the disease, but did not develop it because of the program.

Although such a setting is not the easiest one for evaluators, it makes a particularly useful case study because it shows the real challenge that actual evaluations face often enough. In some ways, aspiring evaluators can find their greatest inspiration not from the cases that are polished like a gemstone, but from the ones that show good, solid work in a messy, uncertain set of circumstances. To repeat—the evaluators here did good work, and their willingness to show the evaluation of a particularly difficult situation is an important gift to students of program evaluation.

As may be obvious, most evaluations that include analyses of costs along with outcomes address programs of substantial scope. The KIDS NOW initiative is the program to implement the 20-year Comprehensive Early Childhood Plan for Kentucky started by the Governor's Early Childhood Task Force beginning in 1999. The goal of KIDS NOW was promoting the healthy development of children in Kentucky through early interventions in four main ways—targeted and universal screening programs and health services, services and other supports for families, a range of approaches to improve the quality of early child care and education, and a commitment to monitor the progress of KIDS NOW and promote it broadly. REACH of Louisville (see Case Study 1) evaluated the program with regard to the fiscal year 2008 timeframe.

One of the important considerations about the analyses of costs and outcomes is that even within a single program, different elements can show very different configurations of important components such as expected time frames for benefits or how easily benefits can be given a monetary value. For this reason, some portions of a program may have very precise estimates of a cost–benefit ratio, whereas others can at best support a fairly broad range of ratios. Further, some decisions such as how overall administrative costs are divided among different programs have no single solution. The cost of administrators who oversee an entire program could be divided among programs on the basis of how many staff work in each program, how many clients are seen, how much area in the building is used for the program, or on other details, which would lead to different results. The point is that the idea of determining how much a given outcome costs may be clear conceptually, but the actual practice is generally much messier.

Meeting Needs

The very broad range of needs that form the basis for the KIDS NOW initiative raise aspects of challenge immediately. The problem of neural tube defects (spina bifida is one form), the consequences of substance abuse by pregnant women, hearing problems among infants, vulnerability to diseases among infants with no health insurance coverage, various conditions among infants that can be detected by early blood tests, and threats to oral health are more than just apples and oranges in terms of the difficulty of addressing them together. They do share the common thread of conditions that can be addressed in a primary or secondary prevention way—intervening before the condition is fully developed or even before there is a problem at all. And this fact is why the plan to avoid very costly conditions is a sound one, financially. But it can be hard enough to show that one such program is effective, let alone such divergent ones. The evaluators took on a heavy challenge.

Implementation

Not only were the interventions forms of prevention where it is difficult to show the details of effects, but the nature of the program was similar to the one in Case Study 5—one where services were being provided to affect infants and young children. Many of those working in such settings value the service to these vulnerable youngsters more than a rigid commitment to data collection, and some missing information is the natural result. The evaluators therefore had limited ability to track the exact details of implementation, and general estimates rather than exact figures were common.

Stakeholders

The staff of the programs, state officials, and the mothers and infants served are some of the obvious stakeholders. Various interested parties, especially politicians with particular viewpoints, however, were stakeholders with agendas, although not necessarily identified overtly. A general summary is that the forces of expectation and demands for sometimes conflicting actions were greater in this case than in most of the others described in this book.

Side Effects

None were noted, especially because established outcomes are difficult enough to document in prevention services.

Improvement Focus

One of the important signs of the commitment to improvement on the part of the evaluation team is their clear recommendation of ways to increase the scope of data collection in future years. Admittedly, even important improvement would not erase all the difficulties the team experienced during this evaluation. But the suggested movement was clearly in the right direction.

Outcomes

The report provided a clear breakdown of costs and benefits for six specific elements of the screenings and health services directed at maternal and child health, all expressed in terms of number of dollars returned for each dollar invested in the project. For the Folic Acid Campaign, there was a return of $3.4 per dollar, a return of $2.3 for the KIDS NOW Plus, $3.0 for the Newborn Metabolic Blood Screening, $4.1 for the Early Hearing Detection and Intervention Program, $5.2 for the Immunization Program for Underinsured Children, and $6.6 for the Oral Health Education and Prevention Program. It is noted that these are conservative estimates. For example, the $3.0 figure for the Blood Screening is based on savings for one year compared to the full cost of the equipment. In subsequent years, savings would continue to accrue, but the equipment would continue to function, so this figure substantially understates the long-term savings. But trying to estimate a reasonable life span for the equipment might lead some readers to believe the savings were being inflated. Still, summing the figures that can be clearly identified among these programs, for a cost of $2,814,384, savings of $9,343,500 were produced, a return of $3.3 for every dollar invested.

Nuances

In many ways, prevention services implicitly address the mechanisms because the services are based on causes identified by some theoretical formulation, rather than observation of a symptom to be treated. It is not surprising that mechanisms are therefore not addressed in any further way.

Summary

As noted, this case study provides an example of addressing costs and outcomes systematically, although it also contributes a particularly realistic example of the difficulties evaluators face. Given the difficulties of documenting implementation with limited records, an implicit point is how essential good record systems are for programs, including but not limited to the evaluations of them. Unfortunately, although evaluators generally know this point very well, convincing program administrators of its value so they address the problem sufficiently at the time that programs begin continues to be a concern.

Source: REACH Evaluation, *Open Arms,* REACH Evaluation, Louisville, 2009.

EVALUATOR PROFILE 12

Margaret Pennington, MSSW—A View from Both Sides

Margaret Pennington is currently (2018) the Director of Policy, Evaluation, and Consultation Services at REACH Evaluation, but for more than 30 years, she worked in state government, the last 14 years of which she worked in the Department for Behavioral Health, Developmental and Intellectual Disabilities, first as the Branch Manager of the Children's Branch, then as Division Director, and then as Commissioner of the Department. In program evaluation terms, this meant she was a key stakeholder requesting evaluations of programs she oversaw. Specifically, during that time, she played a key role in the decriminalization of the involuntary commitment process, the design and implementation of a system of care for children with severe emotional disabilities, and the enhanced involvement of consumers and families in the design and delivery of services. Having the perspectives of being both a consumer and now a producer of evaluations gives her particularly relevant insights into the field.

Some of her own story is about following a relatively common pathway into evaluation. In her MSSW program at the Kent School of Social Work at the University of Louisville, she had courses in statistics and some other formal introductions to general program evaluation points, and she recalls always valuing the process. But even with her highly relevant experience in making use of many evaluation reports, when she began to work at REACH, she found it helpful to spend time learning some of the specific procedures and terminology. Like the vast majority of evaluators, her early formal instruction has been combined with on-the-job learning. Still, she was attracted to REACH because she had seen in their evaluation reports a commitment to conveying the vital information clearly in a manner that made sense to nonresearchers, as well as in a sufficiently brief format that busy people could and would read. Reports like the one on KIDS NOW that form the basis for Case Study 12 show those two important qualities among others.

Margaret made particular note of another common experience among evaluators—that she works as a part of an interdisciplinary team to address the many challenges of program evaluation. She noted the statistician, database developer, epidemiologist, and graphics designer as well as several others along with herself who address multiple issues like qualitative analysis. Although some evaluations of relatively straightforward programs may not demand such a broad set of skills, the challenge of the KIDS NOW evaluation, especially looking retrospectively at services that had important gaps in their records, could be best addressed by a team with many complementary skills. That team clearly functions well in large measure because of effective leadership with a good understanding of the many sides involved in projects.

Source: M. Pennington (personal communication, February 5, 2018).

SUMMARY AND PREVIEW

Cost accounting is a necessary aspect of operating an organization in both the public and private sectors. Efficient use of resources requires that costs be categorized as aids to program selection and management of existing programs. If program benefits can be converted into dollars, ratios of benefits to costs can be calculated. If program outcomes cannot be converted into dollars, programs effecting similar outcomes can be compared to learn which program is most efficient. Even when quantitative indices cannot be calculated, stakeholders must be able to examine outcomes in light of the costs of programs and alternative uses of the funds.

This chapter ends the discussion of evaluation strategies. Experienced evaluators know that the field is an amalgam of quantitative analysis, practical understanding of programs and policies, and skills in working with organizations. These methods and skills produce information for stakeholders; how to report such information and how to encourage the use of information are the final topics in this text.

STUDY QUESTIONS

1. Pretend that you are involved in planning a storefront legal aid program for poor residents. Draw up a budget for the first year. Take educated guesses as to salaries, rent, supplies, and so on. Label each entry as to whether it is a recurring, fixed, hidden, or some other category of cost. Some costs will have more than one label. Group the costs as direct or indirect costs.
2. Compare two reading programs using effectiveness-to-cost ratios. Program A costs $6,000 and results in a one-month reading gain. Program B costs $13,000 and leads to an average gain of a half-month reading achievement. Suppose Program A serves 100 children and Program B serves 300. Which of the two programs is more cost-effective?
3. Compare cost–benefit and cost-effectiveness. Think of the different types of decisions that can be made and the kind of information needed to do the analyses. Summarize your results in two columns.
4. The state of medical technology is such that ill or injured hospital patients can be unable to see, hear, communicate, read, walk, groom, or dress, but yet be considered alive because their hearts are still beating. They may be kept alive through feeding tubes and heart and lung machines. Sometimes extraordinary efforts are made to resuscitate such patients when coronary failures occur. This problem has ethical, legal, and cost considerations. What might a cost-effectiveness analysis suggest about emergency resuscitations for patients who have been kept alive through feeding tubes but with no chance of returning to health or communicating with others? Think about psychological costs and benefits, opportunity costs, and alternative services to other ill patients as you think through your answer.
5. What are the advantages and disadvantages of cost-utility analyses?

ADDITIONAL RESOURCE

Muennig, P., & Bounthavong, M. (2016). *Cost effectiveness analysis in health: A practical approach* (3rd ed.). San Francisco, CA: Jossey-Bass.

This recent edition provides updated details and approaches to this challenging area of analyzing costs and outcomes. Although the focus is on physical health, the principles apply to programs in any area.

13 Evaluation Reports
Interpreting and Communicating Findings

Communication is a multidimensional, interactive process; consequently, to be as effective as possible, evaluation reports must be multidimensional and interactive. While it would be an unusual evaluation without a written summary report, evaluators err if they assume that the end product of an evaluation is a single written report. Without careful consideration of the process of communication, even well-crafted evaluation reports will not be understood, and, when not understood, the information they contain cannot be utilized. Just as we plan controls for threats to internal validity, evaluators need to plan for threats to effective communication.

The first steps in communication were mentioned in Chapter 2. Involving stakeholders in formulating the evaluation questions and planning the evaluation engages stakeholders in the project; this is essential. In large organizations or government agencies in which evaluations are mandated by administrators without consultation with managers or staff, evaluators often seek to work with program staff so that programs are reflected accurately and information is provided that can be used for program improvement. Even if a program director cannot decline to cooperate with an evaluator, the director remains an important stakeholder whose goodwill is helpful. Hegarty and Sporn (1988) described the importance of involving stakeholders in their evaluation work in the U.S. Food and Drug Administration, stressing that effective evaluators avoid appearing to work behind anyone's back. Following the suggestions given in the first chapters improves the chances of getting started in a positive way.

After initial discussions, an explicit communication plan needs to be developed that reflects the information needs of the stakeholders. Next, the principles to guide the preparation of oral and written reports are described. Because few stakeholders are able to draw out implications from tables of numbers, evaluators should make substantial use of graphs and other visual aids. Last, sharing drafts to obtain feedback before finishing the final report is suggested.

DEVELOPING A COMMUNICATION PLAN

Explore Stakeholder Information Needs

The first step in developing a communication plan is to prepare a list of the information needs of the various stakeholders and the times during the evaluation when the evaluators will benefit from feedback from stakeholders. While the stakeholders have information needs, they also provide information. A program evaluator will be more effective if stakeholders can respond to the plans of evaluators and provide interpretative insights that can come only from people intimately familiar with the program.

Plan Reporting Meetings

It is helpful to plan meetings with the stakeholders. Figure 13.1 illustrates a communication plan for a hypothetical evaluation of a new developmental skills program at what can be called Sanders College. The program serves students who have been admitted as part of an outreach program for promising but underprepared high school graduates. The purpose of the formative evaluation is to gather information on

how well the program operates and meets the needs of the students, and to develop suggestions for improvements. Stakeholders include the students served, the staff, the director of the program, and the college dean. Although somewhat removed, other important stakeholders include the parents of the students and the college administration, which approves the funds for the program. Imagine that the evaluation project was planned during the summer and began with student orientation just before fall classes began. The completed evaluation should be available in time for the staff and dean to make changes in the program for the following academic year.

Stakeholder	Information Needed	Format of the Communication	When Information Will Be Available	Setting
Research team	Progress to date, next step in the project	Team meetings, internal memos	Once every 2 weeks	Team meetings
	How to solve problems in the evaluation meetings	Special meetings	As needed	Team and one-on-one meetings
Program director	Progress updates	Oral presentations and one-page memos	Once per month	Meetings of the lead evaluator and director
	Final Report of recommendations	Oral presentation and written final report	March 15	Formal meeting
Staff advisory group	Information to be gathered	Oral presentations and discussions	During planning of the evaluation	Group meeting
	Tentative recommendations	Oral presentations and discussions	During the interpretation phase	Group meeting
Student advisory group	Information to be gathered	Oral presentations and discussions	During planning of the evaluation	Group meeting
	Tentative recommendations	Oral presentations and discussions	During the interpretation phase	Group meeting
Dean of the college	Progress updates	Copies of the one-page memos for files	When memos are given to program director	Inter-office mail
	Final Report on the recommendations	Oral presentation and the written Final Report	April 2	Formal meeting

Figure 13.1 An illustrative communication plan for an evaluation of a program to serve the academic needs of underprepared college students.

As Figure 13.1 shows, the evaluators listed the times when different stakeholders would use information and when it would be helpful to get feedback on tentative findings. Meetings are scheduled with the participants (students in the program). Meeting with the participants is often done during program evaluations, but such meetings would have different functions in different types of programs. In a hospital setting, evaluators would probably interview some patients to understand a program being studied, but few hospital patients spend long periods in the hospital and most do not feel well enough to want to meet with evaluators. On the other hand, the participants in the Sanders College program are involved for at least an academic year; thus groups of them would be available for brief meetings. Note that the content of the meetings would be different for the various stakeholder groups because their information needs are not the same. The dean is to be informed through short progress memos, but the director is scheduled to receive more detailed updates.

Set a Communication Schedule

The dates when reports are to be made to the director and the dean are coordinated with the academic calendar. It would be important for the dean to have recommendations to discuss with the director before the planning for the following academic year has begun. By April 2, many incoming students will have been accepted into the following year's program; however, it is not too late to increase or decrease the number of students involved relative to the first year. Also, there is time to change the level of financial support or to plan some changes in the staffing pattern. Many of the reporting dates could be altered as conditions warrant; however, the dates for reporting to the director and the dean are inflexible because they must have time to

make plans for the following year. In other settings, key dates might be determined by the fiscal year, deadlines for budget preparation, or dates for grant renewal applications.

PERSONAL PRESENTATIONS OF FINDINGS

Evaluators learn quickly that managers are busy and unlikely to spend a great deal of time puzzling over written evaluation reports. Consequently, evaluators prepare reports compatible with the learning style of managers; oral presentations are usually appropriate (Hendricks, 1994). Although the content of these reports is different from written reports, the director and dean should have a written report before the personal presentation. Although the staff may have paged through the report before the meeting, evaluators do not assume that other stakeholders have read it before the meeting.

Need for Personal Presentations

In spite of our high-tech communication systems, people often respond more favorably to person-to-person contact than to other forms of communication for many purposes, although text messages and e-mails conveying focused points such as travel warnings are becoming increasingly common. Still, in a real sense the communication of evaluation findings is a form of attitude-change intervention (see Posavac, 1992). Ideally, evaluations are sought because people are looking for new ways of organizing or directing programs. Learning occurs more easily through interpersonal dialogue than through written material (Preskill, 1994). One reason for this is that people pay more attention to a person than to words on a page or a monitor screen. Personal presentations are also effective because the presentation can be fit to the needs of the audience. Questions and facial expressions showing puzzlement, skepticism, or agreement provide evaluators with information that can be used to improve the presentation.

Content of Personal Presentations

A personal presentation usually includes a short review of the background of the evaluation. In the communication plan illustrated in Figure 13.1, the dean is to receive written progress reports. Consequently, only a brief reminder of the reasons for the evaluation and an overview of the procedures used are given in oral presentations. Most presentation meetings last between 30 and 60 minutes; the time available should be focused on describing the findings and recommendations and fielding questions. The details of the sampling procedure and data analysis would not be discussed because the primary concern of the decision maker (in this case, the dean) is the conclusions. If evaluators have gained the trust of stakeholders, administrators are not going to second-guess the methods used.

Using computer-projected visuals helps to make major points and keep the attention of the audience on the message. Each image should focus on one major idea, such as the impact model, a graph of findings, or recommendations, and provide some, but not too much, explanation. A set of handouts matching the images could be distributed (Hendricks, 1994). See the eResources for examples of appropriate images and handouts as well as those with too little and too much information.

A major misunderstanding, and a devastating experience for evaluators, occurs when stakeholders claim that the findings were already well known and that an evaluation added nothing to their knowledge of the program. Of course, it is true that an evaluator might work independently from the stakeholders and indeed duplicate existing knowledge. If evaluators work closely with stakeholders as recommended in this text, however, it is more likely that some of the information will be new. Claims that "nothing new has been found" may be an example of hindsight bias (Guilbault, Bryant, Brockway, & Posavac, 2004). Researchers have documented a tendency for people to incorporate new information into what they know, and come to believe that they "knew it all the time." Professionals receiving evaluation reports are also susceptible to this bias. This tendency should not be confused with a conscious attempt to deny threatening evaluations or to avoid working with new ideas; hindsight bias seems quite common and automatic. A challenge to the evaluator is to demonstrate to stakeholders that the information is new to them while not damaging the interpersonal relationships between the evaluator and the stakeholders.

Some evaluators seek to get stakeholders to state explicitly what they expect the evaluation to show.

Because people seldom like to make statements publicly that might prove to be incorrect later, this strategy needs to be handled carefully. The following procedure has been used successfully: (1) listeners are given a graph that includes the axes, scales, and a title, but no data; (2) they are asked to plot the results that they expect the evaluation to have found; and (3) then the complete graph with all the data is distributed. When listeners do this, they can compare their own expectations with the data (this minimizes hindsight bias), yet they do not have to reveal publicly that they were wrong. This has worked well in settings in which managers meet to receive feedback on their employees' work attitudes on a variety of scales. Managers are given the scale averages of a whole division, but not those of their own units until after they committed themselves privately to how they think their own employees feel. No one knows how right or wrong any individual manager was; however, each manager can see what information he or she anticipated correctly and what findings are indeed new. Note that this procedure also involves listeners in the feedback process rather than letting them remain passive. With evaluations of small agencies, perhaps evaluators can learn what has been found in evaluations of similar agencies and ask managers to note how they think their agency compares to others.

Audience for the Personal Presentations

Along with the dean of Sanders College, others present for the summary report would probably include the program director, the assistant dean directly responsible for the program, the evaluator/presenter, and an assistant to the evaluator. The size of the program influences the size and nature of the audience. When federal programs are described, a much larger number of people would be present compared to those concerned with the small program at Sanders College. Even with small projects, it is helpful for more than one member of an evaluation team to participate in the presentation. The second member can distribute the handouts, handle visual aids, supply information the presenter may have forgotten, and clarify questions. Since the assistant is not responsible for the presentation, he or she can assess the audience's reaction to the presentation.

Distributing Drafts of Reports

Before the personal presentation to the program director and the dean, a written draft should be available to the director. This version should be called a draft and labeled "Not for distribution." Although the interpretations and recommendations of possible areas of improvement have been presented in conversations with the staff and students, recommendations are subject to change. After looking over the draft report, the director may be able to suggest additional recommendations or may point out errors. A discussion of the draft is likely to result in changes in the report.

Calling this document a "draft" avoids offending the most senior administrator, who may feel that the final report should come to him or her before being released elsewhere. If any concern is raised, the evaluator can point out that drafts have been discussed, but that the final report is being presented first to the administrator. After the oral report and with the administrator's approval, the written final report can be made available to others. Organizations vary in how formal they are regarding handling information; among the many reasons to work closely with stakeholders is to develop a good understanding of the assumed communication patterns in the organization sponsoring the program.

CONTENT OF FORMAL WRITTEN EVALUATION REPORTS

The personal presentations to decision makers are the single most important report. However, there are many other channels for reporting program evaluations; this section describes the content of written reports.

Remember the Purposes of the Formal Report

Evaluators are nearly always expected to submit a report describing a program evaluation. Even when major stakeholders have been regularly informed about the progress of an evaluation and have participated in an oral briefing, a formal report is necessary. First, it is crucial to provide a written record that an evaluation was carried out. Those who arranged to fund the evaluation may need a tangible report for budget accountability.

Second, a written report can answer questions about the findings of an evaluation or about its methodology that may arise later. Third, if administrators make a decision on the basis of an oral briefing, they will want a written report as backup if others question the reasons for the decision. Fourth, when administrators, program staff members, and evaluators take new positions or retire, their memories of the evaluation are lost. A written report remains for others to examine: write with that in mind. Use dates, not phrases such as "last year"; include the individuals' job titles, not their names; and put a date on the cover page. Without a written report, some people may propose a new evaluation of a program already studied. It is not unusual for members of planning committees to suggest gathering data on an issue only to learn that an evaluation dealing with the same question was completed a year or two earlier. Even if planners decide that an additional evaluation is needed, access to the written report of a previous study would be very helpful to future evaluators.

Provide an Outline and a Summary

While most program directors read evaluations of their programs carefully, many administrators read the summaries only, especially the brief reports labelled "Executive Summaries." These summaries are not like abstracts that precede many research articles. Abstracts often contain phrases such as "the implications of findings are discussed" without mentioning the implications at all. Unlike journal abstracts, summaries are longer and can stand alone. A summary is a short version of the report including the recommendations; conclusions are not saved for the discussion section. Summaries of internal evaluation reports are often two pages long although their length varies with the complexity of the evaluation. Evaluations of large government programs or nonprofit organizations will be more ambitious and the summaries will be longer. Still, providing an outline and an Executive Summary are standard parts of reports.

When writing reports, some beginning evaluators are hesitant to use subheadings and instead write page after page providing readers with little assistance in sensing the general flow of the report. Readers of evaluations need all the help we can give them. Think about reading this text with no subheadings in the chapters; without subheadings, a textbook is hard to follow.

Describe the Context of the Evaluation

A written report includes a description of the program setting, the program staff, and the purpose of the evaluation. These topics would be mentioned only briefly in an oral report because the audience often includes those who are quite familiar with the program and may have requested the evaluation; furthermore, they have been kept informed about the project as it was carried out. However, it is impossible to anticipate who will read an evaluation report in the future; consequently, it is written with the assumption that readers know little about the program.

A description of the program that was evaluated is important also because it is impossible to predict accurately how a program will evolve, how funding will change, or how directors will alter it. For future readers to make sense out of the report, the program itself must be described. Some of these descriptions can make use of material already prepared by the program and simply be placed in an appendix. It is often surprisingly difficult to find descriptions of university programs several years after the director of a program has left a university.

Although a majority of evaluations are formative (designed to develop ideas for program improvement), it is essential to describe the purpose of the evaluation in the report. The precise issues that were to be addressed should be listed. There are major differences, for example, between a project to assess the training needs of staff members and a project to evaluate the effectiveness of current training programs, even though the two projects might well measure many of the same variables.

Describe the Program Participants

Reports usually include a description of those who received the services of the program. Demographic information helps others to judge whether the program was particularly suited to the people served and whether the program might be appropriate in other settings. Beyond information such as age, education, and gender, a description of participants might include how the people came to enter the program. It makes a

difference whether the people served represent the target population for whom the program was designed or whether they learned about the service and sought it out on their own and, thus, may be different from the intended target population. The rate of attrition and the descriptions of those who did not complete the program would be helpful to anyone hoping to improve it. Remember a point from Chapter 5: guard the confidentiality of the participants.

Justify the Criteria Selected

Whether the project was a needs assessment, a formative evaluation, or a summative evaluation, certain criteria were selected to indicate unmet needs or program outcomes. Indexes of unmet needs should be related to the problem on which a needs assessment was focused; outcome criteria should be described and related to the objectives of the program. In a report of an outcome evaluation, this section might be an appropriate place to provide the impact model of the program.

Evaluators are coming to appreciate more fully how weaknesses in observation techniques reduce the statistical power of an evaluation to detect an effect of a program (Lipsey, 1990). Evaluators may want to document that they considered these issues when they planned the evaluation; however, it might be a good idea to place technical material regarding the consideration of statistical power and the reliability and validity of measurement tools in an appendix since this material could be difficult for many administrators and service delivery staff.

Describe the Data-Gathering Procedures

It is important to describe procedures used to obtain data because the way information is gathered can affect the findings. Here are three examples: poorly trained interviewers extract less information than well-trained ones; the 2000 Census is less relevant for describing current unmet needs than the 2010 Census; and self-report surveys are subjected to more self-interest biases than material gathered from records. It would be good to show that the people responsible for data gathering were independent of the program evaluated.

Provide the Findings

Use Graphs

The availability of user-friendly software has reduced the need to have graphs drawn professionally for all but the most complicated and expensive evaluations. Figure 13.2 shows the value of displaying not only the means of groups, but the confidence intervals as well (Cumming & Finch, 2005). The innovative treatment group (CRP) showed a better outcome than the traditional program (TP); however, the mean for TP participants lies within the 95% confidence interval of the CRP group. Clearly, this would be weak evidence for the value of the innovation. In the right panel, the means and confidence intervals provide a much more favorable outcome. The confidence intervals provide a way to place the differences among the means into a context. The set of findings (on the left) does not provide confidence that the difference observed could be replicated, but the set of means on the right does permit concluding that the innovation led to a better outcome. Note that the range of possible scores for the outcome measure was given in the vertical axis. This range is a second way to place the findings in a context. In program evaluation, stakeholders want some confidence that program differences have a practical import for participants. Seeing group means in the context of the worst and the best values of the scale helps to interpret the differences found. In basic research, the focus is on finding any difference that would reflect on the theory being tested; in evaluation, it is important to know the level of the outcome, not just relative level. By the way, it should be remembered that just because a type of graph can be made by a computer program does not mean that the graph is a good one. See the eResources for more examples of graphs and other visual displays, such as the stem-and-leaf plot noted below.

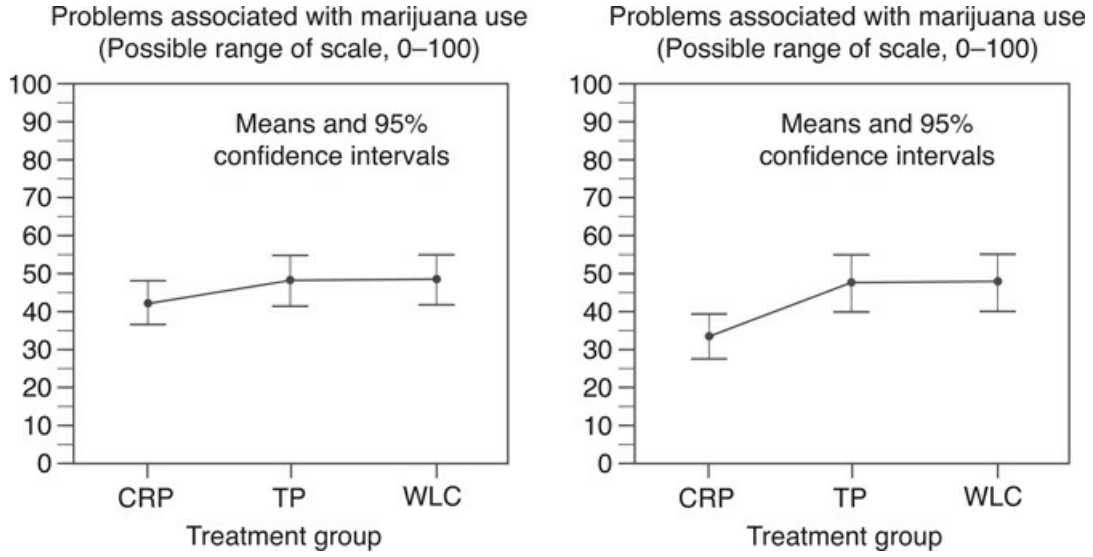


Figure 13.2 Illustration of the value of including 95% confidence intervals when reporting means of groups.

Anticipate Misunderstandings

There are a number of interpretation pitfalls for evaluators and readers of evaluations. While program evaluations share these problems with other research methods, the problems may be especially critical to evaluations because inaccurate or misleading interpretations of evaluation findings can have serious implications. Recognizing the possibility of such misunderstandings encourages evaluators to prepare reports written so as to minimize the possibility of such misuse of findings.

Many audiences, including some well-educated ones, confuse the size of a sample with its representativeness (Nickerson, 2000). Information from large numbers of self-selected people responding to a survey is often overvalued, even when a low response rate should indicate that the findings cannot be treated as describing the population of interest. Smaller, carefully sampled representative groups yield more accurate estimates of the population values, but may be criticized unfairly simply on the basis of size.

Although many people can understand the implications of a difference between two means, standard deviation is a very hard concept for people to understand. Readers sometimes interpret a statistically significant difference between two group means to indicate that the distributions of the scores of the people in the two groups barely overlap (Posavac & Sinacore, 1984). The way to avoid misinterpreting differences between means is to graph data. Suppose that a university department chair becomes concerned about the difference between the grading practices of faculty and teaching assistants (TAs) in her department. One approach to examining this question would simply be to compare the means of percentage of A and B course grades assigned by faculty members versus TAs. Suppose that she learns that the mean percentage is 53% for faculty and 63% for TAs. The difference in the means is statistically significant ($t(116) = 3.723$, $p < 0.0005$). The effect size is 0.70, indicating a large difference relative to the guidelines suggested by the findings of Lipsey and Wilson (1993). Many stakeholders are likely to believe that the low probability reveals that most TAs assign higher grades than most faculty members do. This misunderstanding can be avoided if the difference between means is interpreted in light of variation within groups. Looking at the two distributions of percentage of A or B grades, it is easy to see that both groups of teachers show wide variation in the level of grades assigned. In presenting evaluation findings, anticipating common misinterpretations increases the usefulness of the report. If the dean decided to act on these findings or to seek more information, the wide variation would be important to understand.

Another common misinterpretation is to focus solely on the size of the numerical difference between two means or two percentages. Differences cannot be interpreted without concern for the nature of the variable involved because even small numerical differences might have marked social importance. Most programs, even well-implemented ones, produce only small changes in outcomes. Consequently, interpretations of group differences are best made in the context of the size of program effects actually observed among many programs (Light & Pillemer, 1984). As mentioned earlier, Lipsey and Wilson (1993) gathered nearly 300 summaries of

evaluations that involved thousands of individual evaluations. They found that many programs produced effect sizes of only 0.25. That value indicates that the mean of the program group was only about a quarter of a standard deviation larger than the mean of the comparison, or wait-list group. Even smaller effects have proved important, especially in medical areas when the dependent variable is life or death (Rosenthal, 1990).

Type II errors are overlooked so often (Freiman, Chalmers, Smith, & Kuebler, 1978; Lipsey, 1990; Shadish, 2002) that the issue bears repeating. It may be that the resources provided to the evaluator were insufficient to perform an evaluation with enough statistical sensitivity to detect a socially valuable effect. Even small reductions—one-half of a percent—in drug use and alcohol abuse can have valuable social benefits (Caulkins, Rydell, Everingham, Chiesa, & Bushway, 1999) and yet be impossible to detect without huge samples. Although none of us enjoys pointing out that one of our evaluations may not have been designed well enough to detect a socially useful effect, recognizing this possibility is better than writing a report that criticizes a program rather than acknowledging the limitations of the evaluation. Report readers who interpret "$p < 0.05$" as program success but "$p < 0.07$" as program failure do not understand the high rate of Type II errors (Abelson, 1995).

Use Appendices

The description of findings cannot be done thoroughly in an oral report and can become quite lengthy in the written report. Because the audience of evaluation reports includes many people not schooled in research methods or statistical analyses, details such as tables of means, standard deviations, and probability levels are better placed in an appendix to the report. The value in following this policy is to produce a complete record of the findings for anyone who would like to trace the connections between findings and recommendations. Also, evaluators usually want to demonstrate that they did carry out the analyses required to support the interpretations they have made.

Develop Recommendations

Sometimes evaluators have concentrated their efforts on evaluation design, data collection, and analysis, developing recommendations nearly as an afterthought. Schedules need to be planned so that there is time to work on understanding what the evaluation means and which recommendations make sense for the program (Weiss, 1988). As the communication plan in Figure 13.1 implied, recommendations are best developed in consultation with the stakeholders. Tentative recommendations can be suggested to stakeholders to learn if the recommendations are reasonable in light of limitations on the agency. Depending on the relationship between evaluators and program managers, these negotiations would be carried out in different ways. When evaluators have been working at the request of program managers, the recommendations will be heavily influenced by the agency. Ideally, an evaluation will have provided new insights to the program managers who intend to use the information to improve the agency. Furthermore, the different perspective of the evaluation team may persuade the agency staff to refine their understanding of the target population or the outcomes that are thought feasible; however, evaluators have no power to influence an agency beyond what it is willing to do. By contrast, when the administration of a facility has ordered an evaluation of the programs it sponsors or when a government department commissions an inspector general's office to carry out an evaluation, the evaluation team can make recommendations that the program managers may not immediately support. This does not mean that an evaluator works without consulting the program managers; often initial drafts are shared, as previously mentioned. However, if the evaluation team cannot agree with the program management, there is no reason to change the recommendations. Unresolved disagreements can be indicated in the written report by including a statement from the program manager, as is the practice of the Office of Inspector General of the U.S. Department of Health and Human Resources (Office of Inspector General, 2009).

Recommendations need to be kept within the psychological reach of the stakeholders. Recommendations that would require major changes in the organization will simply be rejected. Even if an evaluation suggests that there are major problems with a program's plans, its implementation, or its effectiveness, evaluators have greater chances of influencing a program when the changes recommended are considered feasible to

implement. If an evaluation was designed to contribute to program improvement, then it must not be thought of as a way to grade the program. As long as no malfeasance or injustices were discovered, the purpose of an evaluation is to support improvement regardless of how well the program functions.

Evaluators are careful to avoid assuming that they are experts in how programs should be designed and managed. Sometimes stakeholders will not want any recommendations. Perhaps the word *recommendations* can be avoided when evaluators believe that an evaluation points to needed changes or new initiatives, but they want to avoid appearing intrusive. Maybe a term such as *implications* or *insights* might be softer, permitting the raising of issues without offending management sensitivities.

Formal Reports Should Look Attractive

Written reports are prepared with care both for content and appearance. In applied settings, managers deal regularly with consultants and salespeople who present attractively prepared proposals and descriptions of new products. Evaluators lose credibility if their reports look like term papers; it may not be fair, but it's true.

Evaluators often place the logo of the agency on the cover of the report. One can use a report binder with a clear plastic cover so the cover page of the report shows through. The most inexpensive laser printer produces what to a nonexpert's eye looks like typeset printing. Using colored paper for appendixes serves two purposes: many people feel that the report is indeed prettier and they are assured that they are expected to read just the white pages—not the whole report. It would be best to choose muted colors, not pink or goldenrod.

PROVIDE PROGRESS REPORTS AND PRESS RELEASES

Since evaluations take at least several months to complete and since there may be some misgivings about what may be learned through an evaluation, critical stakeholders should be informed about the progress of the project from time to time. Progress reports can be made on a regular schedule (as suggested in Figure 13.1) or at milestones in the evaluation cycle. For example, progress reports might be made when data collection begins and when collection is completed. Progress reports can also be used to thank staff members who cooperated in one or another phase of the evaluation.

When the community is a stakeholder, press releases might be appropriate to report to people outside the agency (Torres, Preskill, & Piontek, 2004; W. K. Kellogg Foundation, 2010). Of course, before an evaluator releases information beyond the agency, it is essential to have an agreement with the appropriate stakeholders. Some private agencies may not permit such publicity; others may crave it. By contrast, reports of some government-sponsored evaluations are legally public documents once the report has been accepted by the director of the evaluation office involved.

Although it is beyond the scope of this book to address international program evaluations comprehensively, Case Study 13 summarizes the evaluation of the Afghanistan Agricultural Extension Project II (AAEP II), funded by the United States Agency for International Development (USAID). The 58-page final report (see eResources) is a good example of matching the report to the scope of the project—most smaller projects do not require such extensive detail. Further, while there are substantial common features across all evaluations, this case illustrates how some aspects of many international evaluations differ in degree from most evaluations in the United States. For example, although funding sources are often distinct from those providing the services, such as when health insurance pays for treatment, in many international settings, the donor is in another country than the program. Potential conflicts in expectations such as reporting requirements are thus more likely and often more extreme. Likewise, although change is common in all settings, the constant threat of change is also more likely and often more severe in many international settings. In a more positive light, recent technology sometimes provides rather striking opportunities. In agricultural projects such as AAEP II, where improved practices are at relatively unsophisticated levels, smartphones permit staff to video one farmer demonstrating a technique and then show that video the same day to the farmer's neighbors. Evaluations may thus span a greater range of techniques and technology than in many other settings. Further, with a budget over $20 million and building on Case Study 12, the project provides a good example of how the element of finances can be relevant to an evaluation.

CASE STUDY 13

Afghanistan Agricultural Extension Project II (AAGP II)

From October 1, 2014 to July 31, 2017, USAID improved agricultural production in 27 provinces in Afghanistan in a capacity building approach that included training Afghan public sector extension workers and community volunteer lead farmers to disseminate improved practices to others. This summary can only cover a limited number of the elements of this extensive project, but the main program evaluation components are clear.

Meeting Needs

Although there was a separate, earlier needs assessment (see eResources), the need for improved agricultural practices and productivity was not addressed in explicit detail in this report; however, the circumstances were indirectly addressed, such as noting that one "research farm still had scattered remnants of destroyed military equipment." Other needs can also be inferred from reported outcomes, such as reduction of losses of grain from pests and other problems in the amount of 76,710 kg by using better storage bags.

Implementation

Extensive details are provided regarding the number of training events in several forms such as the Training-of-Trainers workshops and the Farmer Field Demonstrations as well as the mini-funds—small loans for specific needs such as a solar panel well pump. During the project, 1,307 Training-of-Trainers workshops were offered to 8,568 participants. As noted in the report, the cross-cutting theme of gender equality and female empowerment made the detail that 597 of the 1,307 trainings were led for women particularly relevant.

Stakeholders

As with most evaluation reports, stakeholders are implicitly noted rather than explicitly identified as such. Still, various Afghan government agencies and employees are described at various points as are the farmers involved in the project. U.S. taxpayers are also inferred as stakeholders—the USAID logo notes, "From the American People."

Especially relevant to the evaluation, a number of challenges regarding work with the various stakeholders were noted. For example, although the Ministry of Agriculture, Irrigation and Livestock was obviously related to its Directorates of Agriculture, Irrigation, and Livestock, the Memorandum of Understanding (MOU) between USAID and the Ministry did not function smoothly with all Directorates. In the "Lessons Learned" section, the report recommended that future efforts develop MOUs with each Directorate to avoid similar problems.

Side Effects

A number of challenges about the evaluation process were noted such as the difficulty the evaluator had registering the project because of discrepancies between UC Davis policies and standard procedures in the Ministry of Economy in Afghanistan, problems due to increased concerns about security over the course of the project, and the project ending more quickly than planned because of budgetary developments. One positive side effect for women was being included in their households as decision-making leaders, a role normally reserved for men.

Improvement Focus

There appeared to be a clear emphasis on presenting findings in a way that laid the groundwork for improvement, such as providing context and suggestions when identifying problems, barriers, and lower-than-expected success. For example, when a number of extension workers who attended Training-of-Trainers events did not lead any Farmer Field Schools, both they and those who did lead Farmer Field Schools were asked about reasons, which provided useful information for improvement.

Outcomes

The report included extensive sections regarding specifics about the broad range of outcomes—such as noting that crop yields were increased, providing 60 metric tons more wheat, that storage losses were reduced by 20%, which can be translated into enough food to feed more than 75,000 people for a day, and that more than 1,000 people received formal training in improved agricultural practices. Further, specific details about the cost for some outcomes were also provided, such as noting that the investment in a silkworm cooperative involving 140 women showed a 23% return above the amount invested.

Nuances/Mechanisms

The theoretical assumptions regarding some mechanisms for the general outcomes were presented clearly early in the report: (1) workgroups were to be formed, which would identify needs and form plans to be implemented in (2) Provincial Model Teaching Farms, the findings of which would be shared in (3) the Training-of-Trainers workshops. Those trained would then spread those insights through (4) Farmer Field Schools and then through (5) Farmer Field Demonstrations. Explanations of the degree to which this model functioned was presented in a narrative fashion, such as noting that although substantial numbers of those trained in workshops did provide Farmer Field Schools, meaningful numbers of those trained did not lead such schools.

Summary

In the drafting of this final evaluation, the authors made every attempt that they could to put data collected over the life of the program into the proper context so that future programs and even similar programs currently being implemented would benefit from the insights and challenges that the AAEP II built expertise around. On a very personal level, the Director for Monitoring and Evaluation of the AAEP II came away from the final evaluation report writing with the conviction that for development programs to reduce dependency, funds must be piped into the places that most need them and directly to the people who will put them to the best use. She hopes that when a follow-on program is announced, the mini-fund program will be expanded in terms of budget share. It provides a natural incentive that helps to support the theory of the program, providing funds directly to extension workers, local researchers, lead farmers, and community groups to overcome shared barriers to agricultural productivity and profitability. This is an approach that has been shown to be highly effective in other contexts and in sectors beyond agriculture (see USAID's learning lab at https://usaidlearninglab.org/).

Source: University of California at Davis, 2014.

EVALUATOR PROFILE 13

Kim Castelin

It should come as no surprise that Kim was the Director for Monitoring and Evaluation for the capacity building project described in Case Study 13, given her profound commitment to building program evaluation capacity in others rather than trying to reserve such knowledge and skills among an elite group of professionals. She notes clearly, "I love teaching others about how to structure and apply evaluation thinking to programs, so that iterative changes can be made. I believe that anyone can be empowered as an agency for change, and that by including local stakeholders closely into evaluations, the evaluation process itself contributes to their own skills and abilities to critically think through complex issues like climate change and land tenure rights or gender equality and cultural preservation." We evaluators who work in the United States often see examples that support Kim's perspective, but in her work across numerous international settings, the illustrations are, if anything, even clearer.

Few people in international development purposefully pursue career tracks in monitoring and evaluation. Similar to many people who find themselves excelling in that field, Kim stumbled into it through an internship in southern Tajikistan one summer. During that three-month period she worked with field teams to conduct farmer surveys, including recall surveys, for a newer program's baseline data. While standing out of the baking sun next to a canal and under a large willow, watching farmers and her colleagues scratch out percentages in the sand, Kim recognized that all of the information the project needed was readily available, but the questions on the survey weren't asking the right questions. This experience was a crash course in learning alongside her colleagues the benefits of a streamlined, locally centered monitoring and evaluation system. That initial introduction, and the realization that much of evaluation work hinges on asking the right questions, have steered Kim's international development career. Her natural comfort in analyzing and interpreting numbers and her fascination with the human spirit lend her the skills she needs to employ both qualitative and quantitative data.

Kim continues to support monitoring and evaluation work in international development primarily by building and supporting preexisting local talent. International development programs work best when those intimate with local conditions are given space to probe, learn, and advocate for connecting resources to people and places that will put them to the most effective use. She is able to extend this capacity building through an online course as well as through in-person training, support trips, and workshops.

Source: K. Castelin (personal communication, February 12, 2018).

SUMMARY AND PREVIEW

Communicating the findings of an evaluation cannot be treated as an afterthought; communication should be planned when the evaluation is being planned. Communication is more likely to be carried out if a plan has been spelled out explicitly from the start. Evaluators are wise to keep stakeholders apprised of progress, especially when the evaluation takes more than two months to complete. A personal oral presentation of the findings is the best way to present the findings to those who might use the evaluation in making decisions. Evaluators are wise to anticipate misunderstandings and provide appropriate information to minimize them. A summary written report documents the details of the evaluation, serves as the official record that an evaluation did take place, and includes details about the procedure, findings, and statistical analyses. In addition, briefer reports and press releases are useful forms of communication with stakeholders.

When evaluators report their findings, they are concerned especially with encouraging the stakeholders to utilize the findings. Although stakeholders are the ones deciding how to use an evaluation, there are some practices that evaluators can follow to increase the probability that good recommendations are followed. These practices are described in the last chapter.

STUDY QUESTIONS

1 How might a communication plan for a qualitative evaluation differ from one for a quantitative evaluation?
2 Suppose that you had conducted an assessment of the needs of student athletes for academic assistance. Think of the variables you might gather. Develop a form that could be used to get stakeholders to make their expectations explicit before receiving the findings of this needs assessment.
3 Find a journal article describing an evaluation. Prepare an outline of an oral presentation of the findings of the evaluation. Prepare an outline of a written report for that evaluation. Note how these outlines differ from the outline of the original journal article.
4 Prepare a summary of the article you have been using. Note how the summary differs from the abstract.
5 List the stakeholder groups who should be part of the feedback process for the report you have been using. What are the differences in the type and detail of the information these groups would want to know about?

ADDITIONAL RESOURCE

Torres, R. T., Preskill, H. S., & Piontek, M. E. (2004). *Evaluation strategies for communicating and reporting: Enhancing learning in organizations* (Rev. ed.). Thousand Oaks, CA: Sage Publications.

This still-valuable volume contains an incredible amount of terrific, practical advice. Beyond advice, however, the authors provide many illustrations of tables, graphics, charts, and handouts. This work shows how far evaluators have come in developing communication strategies that go far beyond what was standard practice not many years ago. The major lesson as emphasized in the current chapter is that effective evaluators plan for communication. While conducting an evaluation, it can be very easy to focus on the technical issues of getting, analyzing, and interpreting data, forgetting that all that work can only be useful if others understand the findings. This work will help you.

14 How to Encourage Utilization

Presenting reports in a way that gets attention is but one aspect of encouraging stakeholders to use evaluations. Unfortunately, in many evaluations a report has been the only effort to encourage utilization. This chapter examines obstacles to utilization so that evaluators can recognize and perhaps deal proactively with them. Evaluators can contribute to an organization if they can assist in developing a learning culture in which stakeholders are willing to consider the implications of the findings of evaluations. Even when the interpretations of evaluations are ambiguous, evaluations often have implications that might not be evident from a superficial examination of the findings.

A number of evaluators have studied how evaluations are used. Here are a few examples. Leviton and Boruch (1983) followed up on 21 federally sponsored evaluations of education policy. They found clear evidence that the evaluations had affected the programs evaluated and the development of new laws. Patton (2008) described various examinations of the effects of health program evaluations; he reported that evaluation information slowly became part of the knowledge base and reduced the uncertainty of administrators. After the introduction of a management information system (see Chapter 7), the managers of a large, residential agency for children discovered that some widely accepted indicators of good service to children did not apply to their agency. As a result of information about the children served, the managers changed some of their practices (Overstreet, Grasso, & Epstein, 1993). Many evaluations are utilized, but evaluators have to recognize the complex role that agency managers fill and the multiple uses of evaluations (Weiss, 1998). It may be, however, that some use of evaluation has been premature because federal procedures for labeling programs successful have been flawed (Weiss, Murphy-Graham, Petrosini, & Gandhi, 2008).

OBSTACLES TO EFFECTIVE UTILIZATION

There are several common obstacles to implementing findings from evaluations. Evaluators are not always in a position to remove these roadblocks, but if they are aware of them, they might be able to minimize their effects.

Constraints on Managers

Program managers may not be able to utilize findings from evaluations because there are many influences that restrict their freedom (Sonnichsen, 1994). It may seem to evaluators that decision makers do not pay sufficient attention to evaluation findings and that they do not wish to reexamine policies in view of the information provided. It is just as likely, however, that a manager must satisfy a variety of competing constituencies. Meeting the needs of the diverse stakeholders of a complex agency requires compromises. At times evaluators have overestimated the freedom administrators have in using new information to make program changes; organizational constraints may make it appear that evaluations have been ignored.

Value Conflicts among Stakeholders

When stakeholder groups are polarized on an issue, it is likely that evaluation findings will be hard to use. When people hold hardened positions linked to values, evaluation findings may fall on deaf ears or parties may endorse some parts of evaluations while rejecting other parts. Sommer (1990) found a similar situation when he conducted a study for a local farmers' market. Between the time he began and the time he finished, the board of directors had changed. Because the philosophy of the new board was quite different from that of

the old one, the report was seen as irrelevant.

When stakeholders hold strong views, evaluators are probably helpless to effect much change. When less polarization is found, it may help to seek agreement ahead of time on what would constitute support for certain conclusions. There is no way an evaluator can stop stakeholders from changing their minds once the findings are available; however, discussing evidence in the abstract before information has been gathered may be a way to diffuse some of the effects of strong opinions and encourage people to approach an evaluation with an open mind.

Misapplied Methodology

Evaluators have sometimes tried to accomplish more than their resources permitted. In their review of published evaluations, Lipsey, Crosse, Dunkle, Pollard, and Stobart (1985) showed that although many outcome evaluations were planned to investigate the causal effects of a program on an outcome, relatively few were implemented in ways that would have permitted the evaluation to isolate causal relationships. This can occur when too few program participants are observed, when threats to internal validity are not controlled satisfactorily, or when the outcome variables are not measured reliably or validly as mentioned in Chapter 10.

Limits on resources in the context of overly ambitious or vague evaluation requests make valid experimental or quasi-experimental evaluations impossible (see Moskowitz, 1993). Frequently, people untrained in evaluation methods do not appreciate the challenge of conducting valid research in service settings. A program evaluation that cannot be interpreted damages the reputation of the discipline and wastes agency resources. It is important to match the evaluation design to the needs of the agency seeking the evaluation. Often evaluators can help stakeholders by developing a method to track and verify implementation or by helping the staff to clarify their implicit impact model (Lipsey & Cordray, 2000).

Evaluating a Program at Arm's Length

This text has encouraged evaluators to work closely with stakeholders in planning and implementing program evaluations. This is advisable so that evaluations reflect accurately both the program and its context. Working at a distance without close contact with stakeholders who administer and provide the services will be an obstacle to full utilization when an evaluation is completed. When people are involved in an evaluation, utilization of recommendations to effect program improvement is more likely (Fross, Cracknell, & Samset, 1994; Torres, Preskill, & Piontek, 2004). When an evaluation has been crafted to reflect the staff's information needs, the staff members are going to be more interested in the findings. Stakeholders are far more likely to give recommendations a respectful hearing if they feel the evaluation team understood the population served, the agency housing the program, the program itself, and the staff. Such understanding cannot be gained through meetings with higher administrators and a careful examination of official program descriptions; it can only be gained through time-consuming interactions involving lots of listening to the stakeholders closest to the participants and day-to-day operations.

Ignoring What Is Going Really Well

When evaluators fail to appreciate explicitly what is going particularly well in a program, whether that refers to implementation, outcome, or efficiency, stakeholders may fear that the program has not been well understood. One of the strengths of the appreciative inquiry approach to evaluation is to stress what makes staff and managers particularly pleased about the program (Preskill & Coghlan, 2003). Building on strengths is usually a wise approach to program improvement. An evaluator may feel that his/her charge is to find ways in which a program can be improved and then emphasize what might be changed in order to better implement a program to achieve more favorable outcomes, or to offer a program more efficiently. These issues are indeed essential, but not to the extent of ignoring what is being done well. Noting that there is a good foundation that could be improved by some small additions is generally received much better than emphasizing what should be changed and minimizing the current achievements.

DEALING WITH MIXED FINDINGS

When some findings of an evaluation support a program and others do not, proponents of a program focus attention on supportive findings while explaining away the negative ones. The problem of dealing with mixed results is accentuated when there is disagreement within the agency over the value of the program or when jobs are involved.

Don't Abdicate Your Responsibility

Evaluators could report the findings and leave the interpretation to the stakeholders—to administrators, staff, or funding agencies. This can be accomplished by omitting interpretations of the findings and drawing no implications from the evaluation. There are times when evaluators are asked to carry out a study and explicitly told not to make recommendations. In such cases, the agency has hired an evaluation team for its research and statistical expertise, not for its possible contribution to organizational development. In such cases, allowing others to decide what the mixed findings mean is appropriate.

Usually, however, evaluators can play a meaningful role in program planning and refinement. In addition, program personnel expect evaluators to try to figure out what the mixed findings mean for the program. Evaluators do not act as though the numbers speak for themselves, because, indeed, numbers do not. Findings will be interpreted—if not by an evaluator, then by others with definite self-interests. If evaluators refrain from making clear interpretations, they are avoiding responsibility for what they were hired to do. Although service staff members are seldom trained to translate evaluation or research findings into policy implications, discussions with them will help evaluators understand the implications of the findings, but it is the responsibility of the evaluators to seek to discover them. Some good approaches are using the term "maybe," describing multiple options, and acknowledging trade-offs.

Don't Take the Easy Way Out

There is always a temptation to please the evaluation sponsor. An evaluator working for the agency sponsoring the program might be tempted to place great importance on consumer satisfaction, which is often surprisingly favorable even when more objective outcomes show little program impact. The course ratings of even poor college teachers are usually in the satisfactory range, the vast majority of hospital patients are satisfied with their treatment, and clients typically report being quite satisfied with psychotherapy (Garland, Haine, & Boxmeyer, 2007) and medical care (Press Ganey Associates, 2009). Other bad reasons for not dealing with the unfavorable parts of mixed findings include (1) noting that the outcome measures were unreliable and thus insufficiently sensitive to detect program effects, (2) questioning the size or representativeness of the sample, and (3) suggesting that the program may not have been at full strength when evaluated. Since no evaluation is perfect, it does not take that much creativity to find faults with an outcome evaluation that one wants to ignore. These issues should have been dealt with during the planning phase, not during the interpretation phase of an evaluation.

Choosing to emphasize the more favorable findings certainly may be appropriate when the unfavorable findings concern only less central objectives. However, stressing the positive and ignoring the negative in order to please the evaluation sponsors, to avoid hurting the feelings of friends, or to improve opportunities for continued funding would be unethical. When a program or a policy does not show appropriate positive effects, some other program might use the resources more effectively.

Show How to Use the Evaluation to Improve the Program

The best way to deal with mixed results is to emphasize that the purpose of the evaluation was to help managers and staff improve or refine their conceptualization of the program. A number of evaluators suggest that utilization is often viewed simplistically because evaluators have not thought through what might be utilized from an evaluation (Crompton, 2009; Weiss, 1998). It is simply not true that the only utilization of an outcome study is to continue a good program or discontinue an ineffective one. By reflecting on alternatives with the stakeholders, evaluators can develop more sophisticated applications and recommendations.

To reduce worry among managers and staff about less than totally favorable evaluation findings, an analogy

from medicine might be helpful. Managers are asked to view evaluators in the same way as physicians view laboratory personnel in hospitals. Physicians are responsible for developing medical care plans to address their patients' illnesses. However, before developing care plans, physicians order laboratory tests, x-rays, and other diagnostic procedures. After studying the results of these tests, physicians develop a better understanding of the strengths of their patients and form hypotheses about the causes of their symptoms and the best course of treatment.

After the treatment begins, patients are tested again to evaluate the effectiveness of the medical care. If the results are favorable, physicians conclude that the diagnosis was correct, and treatment is effective; treatment continues. If the test results show that the patient is not improving, physicians conclude that the problem is difficult to diagnose or does not respond quickly to treatment. Alternate treatment plans may have to be developed. Follow-up testing could also reveal negative side effects that must be taken into account as the treatment plan is modified. Although a physician may not be pleased with the test results, helping the patient is facilitated by having the test findings. In a similar manner, managers and staff can work with evaluators to examine evaluation findings that will reveal the strengths and weaknesses of a program.

USING EVALUATIONS WHEN AN INNOVATIVE PROGRAM SEEMS NO BETTER THAN OTHER TREATMENTS

When evaluators conclude that treatment and comparison groups do not differ on outcome criteria, they may face challenges in working with stakeholders. But first, evaluators must be sure of their conclusions. In some cases, there might be a silver lining.

When Can Evaluators Be Sure Groups Do Not Differ?

Evaluators carefully review their own procedures before concluding that a program was ineffective. As mentioned in Chapter 11, a naive use of statistics had led evaluators to believe that an evaluation showed no impact when, in fact, the evaluation design and analyses were inadequate to detect valuable changes in program participants (e.g., Julnes & Mohr, 1989; Lipsey, 1990; Meehl, 1990; Yeaton & Sechrest, 1986, 1987).

Although often ignored in practice, it is widely known that a nonsignificant statistical test can be produced by small samples, by marginal implementation, by unreliable measures, or by demanding that only a very low probability of a Type I error be tolerated. Before concluding that the treatment group and control group did not differ, it is necessary to analyze each of these problems. Ways to compute statistical power, described in Chapter 11 and available in the eResources, can be used to learn the size of a sample that is needed to detect a specified effect size if the program causes it. Thus, the first step is to be sure that the sample sizes were sufficient. In addition, evaluations should include an examination of the degree of implementation. In the early stages of many programs, implementation is marginal and many programs are never fully implemented (Shadish, 2002). Before concluding that a program failed, it would be important to know just what it was that failed. First, was it the program as planned, or a program that included only half of the planned components? Second, were the measurement tools sufficient to capture the essence of the program? Third, some evaluators seem to believe that credibility is enhanced when they test null hypotheses with the same low probabilities as used in basic laboratory research. Holding to a rigorous barrier to rejecting null hypotheses means that the chances of missing a program effect will be large. Here are three alternative approaches to avoiding Type II errors (i.e., missing a program effect). One method is to raise alpha level above 0.05. A higher alpha means the Type I errors will be larger, but Type II errors are smaller. An alpha level might be chosen so that the probabilities of Type I and Type II errors are equal (Posavac, 1998). One may conclude that an innovative program is equivalent to no (or the standard) treatment only after having made a good effort to find a difference (Fitch, 1995). It is important to demonstrate that the probability of a Type II error is very low before saying, "There are no differences between the groups."

Another aspect of making a good effort to find an effect before concluding that a program is just as effective as the control program is to examine the impact model. A useful impact model would have included intermediate outcomes as well as final outcomes and specified characteristics of the participants most likely to benefit from the program. A good effort would have involved testing implications of the impact model: were

intermediate effects found and were correlations between outcomes and participant characteristics observed as expected? Such theoretical questions are part of a good effort at finding a program effect.

Are Evaluations Valuable Even When No Advantage for the Innovation Is Found?

If a good effort has been made, one may conclude that the evidence strongly suggests that the innovation was not as effective as hoped. Negative findings are disappointing; however, there might be a bright side to finding that a standard treatment was as effective as an innovative program. If the control group received the standard treatment, it might be true that a less expensive program is just as effective as a more expensive one. In criminal justice, a critical outcome variable is recidivism. If, for example, a work-release program for prisoners did not lead to lower recidivism of convicts as was hoped, one might think it failed. But if work-release participants were not rearrested at a greater rate, it might still be called a success if it were less expensive to permit a prisoner to participate in a work-release program than to simply keep him or her in a traditional prison. In other words, work-release programs can be supported—not because they are better at rehabilitating prisoners, but because they are cheaper and yield results as good (or as bad) as regular prisons. In such a case, a work-release program could be supported, but for a different reason than suggested initially.

DEVELOPING A LEARNING CULTURE

Some individuals and organizations appear to defensively reject information and suggestions to initiate different activities (Gill, 2010). Change for change's sake is not needed in human service organizations; however, willingness to consider new alternatives is needed to assure that high-quality work is carried out efficiently (see Argyris, 2004).

Work with Stakeholders

The utilization of evaluation findings is more likely if evaluators have worked closely with the staff and managers in planning and carrying out an evaluation and if they continue to work with the stakeholders after an evaluation is officially completed. This approach extends the evaluator's role beyond what had characterized the field in the past; often evaluators were called upon to gather data, present a report, and move on. While evaluations conducted in such a manner could be useful, the probability of effective application increases if the evaluator's role includes working with stakeholders while they implement changes (Fleischer & Christie, 2009). If evaluators continue to gather data and provide feedback about the effectiveness of changes, the cycle of planning, implementing, evaluating, and making program adjustments is completed as suggested in Figure 1.1.

Adopt Developmental Interpretations

The recognition that not every innovation or creative idea will be a success is an essential element of a culture that facilitates effective utilization of evaluation findings. Program directors and staff must feel free to innovate, but they must also be willing to recognize and learn from failures. The willingness to set aside less productive ideas and procedures enables staff to experiment with new approaches to achieve program goals.

When failures are hidden, it is impossible to learn from them. If approaches that failed are unknown, then marginally effective efforts to change organizations go unquestioned and are not improved. A more fruitful approach is to measure competence not only by skill in avoiding errors but by skill in detecting them and acting on the information openly. The attitude of nursing managers toward nurses who reported "avoidable drug errors" (ADEs) was related to how many were reported. Managers who defined their roles to include coaching and the development of their staff were more likely to be informed of errors. Those who used a punitive management style learned of fewer errors and thus had less opportunity to change procedures to reduce future ADEs (Edmondson, 1996, 2004). We must learn to reward those who own up to failures and learn from their errors so that changes in programs can be made. Evaluation procedures should not inadvertently encourage hiding errors.

While an atmosphere that respects freedom to fail is essential for an effective use of evaluation findings, it

is beyond the power of evaluators to singlehandedly create such a culture if it does not exist. In healthy organizations, administrators understand that innovation entails the possibility that even well-designed and fully implemented programs might not work as well as anticipated. When a healthy attitude toward innovation exists, evaluators make every effort to encourage it. It is crucial that evaluators avoid implying that evaluation has a punitive role. It is important to keep staff focused on the point of evaluation—data-driven program improvement—not the preparation of summative report cards. When central administrators request an evaluation of a program, it may be useful for evaluators to request an explicit statement from administrators that the purpose of the evaluation is to encourage reflection and further development, not to expose failure. Of course, it may well be that the least confident administrator will not sponsor evaluations in the first place.

Frame Findings in Terms of Improvements

A third element of a productive attitude involves leading staff and managers to expect that evaluations will include some recommendations of new ways to deliver a service or refinements in the objectives of the program, rather than a stark presentation of good news or bad news (Torres, Preskill, & Piontek, 2004). It is extremely rare that an existing program is terminated or continued solely on the basis of an evaluation. Nevertheless, program managers who are deeply involved with their programs worry about continued agency support. As a result, they often look for one-line summaries of evaluations: does the evaluation support the program or doesn't it? Might this approach be similar to that of students who check the grade on a paper, but fail to read the professor's comments? Evaluators raise the chances of findings being utilized effectively when they have helped program personnel understand that nearly every evaluation brings both good and bad news, and that the important question is, "What directions for program improvement are implied by the evaluation findings?" The goal is to improve and modify programs, not to enshrine or destroy them.

During an evaluation of a physical medicine and rehabilitation unit in a community hospital, the unit team understood from the beginning that the rehabilitation unit was going to remain intact regardless of the outcome of the evaluation. However, administrators were interested in learning how much progress stroke patients were making during their stay in the unit, and also whether those who had improved during the hospitalization continued to improve, or at least maintained their capabilities, after they had returned home. The findings of the evaluation showed that during the hospitalization, stroke patients improved their physical skills (walking, eating, dressing) more than their cognitive skills (speaking, remembering). The evaluation also showed that the majority of patients who made progress during hospitalization maintained their improvement after returning home. This information helped the rehabilitation teams consider whether a follow-up program geared to patients who did not maintain their improvement was needed.

Treat Findings as Tentative Indicators, Not Final Answers

Treating evaluation findings as tentative indicators will contribute to maintaining a learning culture. It would be convenient if one evaluation could provide the definitive word on an issue; however, those who expect a final conclusion from an evaluation are bound to be disappointed (Figueredo, 1993; Shadish, 1993). One reason for treating findings as tentative is that specific community characteristics can facilitate a program working well in one location or with people having particular needs but not in another community or with another population. Although we evaluate programs and develop avenues for pursuing improvements, we cannot expect to generalize about other settings or about what would happen if conditions changed markedly in the current setting. Our findings can be treated as hypotheses about what works best.

Recognize Service Providers' Needs

Evaluators have often focused on research methods and provided valid information without understanding the needs of service providers. One way to avoid this problem is for evaluators to assume multiple roles within an agency (Goering & Wasylenki, 1993). Some evaluators have been trained primarily as service delivery professionals. Those people may have an advantage in being able to conduct evaluations that are carefully targeted to the information needs of service providers and are more likely to be utilized. Besides having a good understanding of the details of agency services, evaluators who provide services themselves have more

credibility with other service providers than methodology-focused evaluators have (Champney, 1993). Of course, a close identification with service providers might limit the likelihood of such evaluators maintaining the necessary openness to the need for change. The best evaluators work closely with stakeholders and learn their professional language so as to recognize their needs and, ultimately, increase the utilization of findings, yet maintain an objective posture toward the program.

Keep Evaluation Findings on the Agency's Agenda

Utilization is more likely when evaluators keep reminding managers and program planners about the findings of an evaluation (Sonnichsen, 2000). It may seem that once an evaluation report has been made, the information presented will be utilized soon or not at all. In fact, often there are times when evaluation information is immediately relevant and times when it is not. Unfortunately, it would not be surprising to discover that an evaluation had been forgotten later when an occasion arose in which the findings could be used. Evaluators would be wise to seek to keep the evaluation on the agenda; someone must remind administrators about the evaluation or it will be forgotten before it becomes relevant. If some readers feel that these comments reflect a cynical view of managers and program staff, they are misreading the point. It is easy to remember something that fits in with our experiences and immediate needs. Unless new information is associated with other experiences, we quickly forget. Consequently, information in evaluations is more likely to be used if it can be related to problems as they become known or to plans as they are made. The best person to watch for application opportunities may be an internal evaluator.

THE EVALUATION ATTITUDE

This text has sought to provide the groundwork for the development of the technical skills necessary to conduct program evaluations in human service, educational, and business settings. Technical skills are only half of the matter, however. Throughout the text, the importance of interpersonal relations and the attitude of evaluators as they work with stakeholders has been stressed (King & Stevahn, 2013). Here several themes are drawn together as a way to highlight these nontechnical issues.

Humility Won't Hurt

Evaluators work in agency settings that are generally not designed to facilitate research. Furthermore, the agency may well have been operating without the help of an evaluator for quite a while. Consequently, evaluators can expect to be seen as providing an optional service. Because we are always working on someone else's turf, any arrogance on our part may nullify the value of superb technical skills.

Impatience May Lead to Disappointment

Program and facility administrators have many constituencies, all expecting and some demanding attention. Program evaluation is only one source of information on which decisions and plans are based. Financial concerns, community pressures, political realities, and bureaucratic inertia are all powerful influences on program planners and administrators. Evaluators with patience are less likely to feel ignored and unloved when their recommendations do not receive immediate attention.

Recognize the Importance of the Evaluator's Perspective

Therapists, teachers, and physicians focus on individuals and often do not reflect on overall programs. Administrators and government officials see an overall perspective, but tend to focus on financial and other practical matters. Evaluators must therefore complement these viewpoints by integrating verification of service delivery, program impact, and resources expended into one package.

Work on Practical Questions

People working in service organizations are seldom concerned about matters of only theoretical interest.

Evaluators can work more effectively when they are oriented toward practical questions about the program being evaluated. Try to remember that helping people change is hard work. Faced with pressing human needs on one hand and with allegations of inefficiency on the other, agency staff members desire practical assistance.

Work on Feasible Issues

If an evaluation is not feasible, do not attempt it. Do not waste your time and that of the program staff. However, although it may not be feasible to work on, say, a cost–benefit analysis, or nailing down clear causal connections, perhaps monitoring activities and outcomes might be feasible and useful to an agency.

Avoid Data Addiction

It is tempting to seek to gather a great amount of information about the program and those served. Asking for too much information increases the proportion of program participants who will provide no information. And, in fact, evaluators often find themselves under too much time pressure to analyze all the data gathered, thus wasting the efforts of the cooperative participants. If the information is not essential, do not gather it.

Make Reports Accessible

Using social science jargon and presenting advanced analyses reduce the flow of information to those not trained as evaluators have been. New evaluators often don't realize that few people understand statistical analyses. Save presentations that display the depth of your knowledge and technical skills for professional meetings. To be effective, presentations must be understandable to those who may act on the recommendations. If findings are not expressed in ordinary language, they should be rewritten.

Seek Evaluations of Your Own Work

When evaluation findings and recommendations seem to fall on deaf ears, evaluators will benefit by asking themselves: Was my presentation clear? Did I address the right questions? Were my answers correct? At times the honest answer will be, "Well, I made some mistakes this time." Seek to learn something for the next evaluation. To evaluate an evaluation, one might use Scriven's (2007) Key Evaluation Checklist or review the *Guiding Principles for Evaluation* (American Evaluation Association, 2004). Having another evaluator provide feedback on the completed evaluation might suggest ways to improve the next one.

Encourage the Development of a Learning Culture

Ideally, evaluation encourages honest relations among staff, clients, administrators, and program sponsors. Failures cannot always be avoided, yet people sometimes act as though failures are shameful. Only the dead make no mistakes; the living fail from time to time. Instead of hiding or condemning failures, help staff and managers to treat honest attempts as experiments and to learn from them.

CASE STUDY 14

Engaging Stakeholders

Dr. Stacy Deck and a graduate student assistant conducted an evaluation of a Financial Empowerment Pilot Program run by the Metro United Way in Louisville, Kentucky that provides a good example of a successful approach to encouraging utilization. One important element was that the evaluation included four separate reports over a 19-month time period, rather than only a final summary. A related point is that the first two reports presented explicitly formative evaluations that not only addressed implementation issues but also suggested modifications that set the stage for a collaborative approach between the evaluator and the program staff through the rest of the project.

The financial empowerment pilot program began in early 2014 (building on elements of the previous program described in Case Study 5). The program provided client-centered coaching services to help low income people set and make progress on goals related to their finances and facilitated seamless access to services related to workforce development, financial counseling, and benefit access. "Program components included service brokerage, an 'opportunity fund' to help families remove barriers to economic success, and peer learning and support" (Deck & Thompson, 2014, p. 1).

Early in the evaluation process, Dr. Deck noted some areas where the most appropriate outcome measures for the intervention would only be possible with some changes in the program. Matching a common component of formative evaluations and engaging the stakeholders in a way that increased the likelihood that the evaluation would be used, she addressed these points and suggestions in the first report only three months into the project. The other reports, especially the final one about one and a half years later, document how effective this approach was.

Meeting Needs

Although program evaluations often focus on the level and extent of the need the program is designed to address, Dr. Deck noted in multiple reports a somewhat unexpected discovery. Most of those served by the program did not fit the stereotype of "poor" people who had low incomes exacerbated by frequent bad decisions such as "predatory or high cost financial products/providers" (Deck & Thompson, 2014, p. 5), fueled by misunderstandings and negative attitudes. Instead the program participants reported quite positive attitudes and practices, so an assumption that clients needed to learn completely new insights and actions was not accurate. This evaluation addressed needs partly by informing the program about the actual needs of their clients.

Implementation

Both early (formative) and final reports documented many details regarding the implementation of the program. This included describing the number and nature of the goals clients established, which ranged from focusing immediate attention on increasing income or decreasing expenses to longer-term strategies such as gaining advanced education. In the first report, it was noted that there were fewer clients enrolled than originally planned, but that it appeared the target would be reached with relatively minor delay. There were also multiple tables documenting relevant characteristics of the clients including various details about their financial status (e.g., how many were above and below federal poverty cutoffs).

Stakeholders

Like most evaluation reports, stakeholders are mostly implicitly noted in this evaluation. However, the evaluation design also explicitly engaged a variety of stakeholders. Quantitative data were collected by front-line staff and from program participants via surveys. Early in the first program year, participant responses to the program and ideas for improvement were elicited during a dinner and focus group. Each interim report was accompanied by a face-to-face meeting with program staff to discuss results and plan next steps.

Side Effects

In order to quantify progress on participant goals and aggregate this information at the program level, Dr. Deck developed a goal planning form for the program. With their financial empowerment coach, participants rated the amount of effort for each action step, and then rated their progress at each coaching session to produce a weighted overall progress rating for each goal. While the initial purpose of this tool was to quantify results for the evaluation, the process of completing it had unexpected direct utility for the program. Participants and their coach found that completing the form led them to deeper discussion of what steps were needed to achieve goals, how hard they would be to accomplish, what obstacles could be anticipated, and how realistic the goals actually were. Seeing the "progress score" increase over time was also an encouraging affirmation of hard work.

Improvement Focus

As noted in the introduction, Dr. Deck clearly emphasized improvement through several appropriate suggestions. For example, she could have selected only outcome measures that fit the financial focus of the program in direct ways. Such measures were part of the evaluation, but as she explored the goals the clients chose as a part of the coaching component, she noted aspects of the process that ran the risk of misalignment. With the focus on client-centered coaching, clients could pick unrealistic goals that they had little chance of meeting. Likewise there was no systematic review of the fit between long-range goals and week-to-week action plans. The recommendation to adapt the client-centered coaching by the addition of these reviews not only made helpful adjustments, but engaged the agency staff more in the evaluation itself and in the use of the insights provided.

Outcomes

The addition of processes to tighten the relevance of goals and also a check that they be attainable were helpful suggestions to improve the program and added helpful outcome measures. Already good aspects were assessed in fairly direct ways—attitudes about budgeting as opposed to unwillingness to plan, knowledge about more productive strategies such as avoiding high-interest options, and standard bottom-line elements like income and savings. The moderate improvements on both goal attainment and improved objective financial standing were described clearly in substantial detail. Among other points, Dr. Deck provided multiple descriptions of elements like income, presenting both group averages and accounts of cases in various categories, such as noting the four clients whose income decreased, which was not obvious in the overall average increase. Further, as with almost every program, some of those who started did not continue. Comparisons between "graduates" and "dropouts" provided additional evidence of the effects of the program.

Nuances/Mechanisms

The main details about the kinds of goals clients set and the addition of the careful process of quantifying relative effort clearly suggested that having a coach help clients set goals and persist in working toward them was a likely mechanism for the positive outcomes. One possible further step would have been more sophisticated statistical analyses to test whether there was evidence to support the idea that the work on goals caused the good results. Among other things, the modest number of participants and the limited statistical power

they provided meant that such analyses were unlikely to be helpful.

Summary

As a clear example that there are often more options available than the extremes, in this evaluation, Dr. Deck built on her discovery that the coaching process was not working as well as intended. Rather than ignoring the problem or just reporting that it was not working well, she pursued a different approach of helping the program staff adapt their way of responding to clients' choices of goals. Such approaches are an important alternative to simply delivering reports - not only do they provide specific suggestions that clients might use, the process of working with clients builds trust, and that greater trust increases the likelihood that the client will make productive use of the information provided. Although cases like this one usually take more time and effort, they are generally much more useful evaluations. As an extra benefit, evaluators are more likely to have a justified sense of accomplishment.

Sources: S. Deck (personal communication, January 4, 2018); Deck & Thompson (2014).

EVALUATOR PROFILE 14

Stacy Deck, PhD—An Evaluator Teaching Evaluation

For Dr. Stacy Deck, program evaluation is not an additional element to be tacked onto her identity as a social worker, but an essential component of her integrated approach to "the covenant of work." She notes that very early in her training, she asked, "How do we know that?" And she came to see evaluation as a responsibility implicit in her client-centered approach and a form of accountability and commitment to high-quality service. Not surprisingly, it appears to be a family trait. Her father, who in retirement volunteered at a food bank, helped the food bank improve their quantitative records of service by adding "pounds of food provided" to "number of people served." This helped the agency to garner added funding, which enabled them to expand their service.

Some of the specific elements of program evaluation have their roots for Dr. Deck in the research course she took as part of her MSW coursework. As she looked for evidence that an intervention was working, she asked more questions: What evidence is there of enduring results? What is the source of these numbers? What percent improvement is good enough? When a practicum placement led to employment following graduation, she wanted to help the agency evaluate their services, but realized she needed still more knowledge and skills. Two former professors provided some needed assistance, and Stacy committed to "paying it forward" in the future, which is evident in the way she goes above and beyond in her teaching now.

The theme of integration is clear in her classes. MSW students learn evaluation by doing meaningful pro bono evaluations, such as interviewing those served at a men's homeless shelter. When one student was unable to complete data collection interviews, Stacy adapted the evaluation design to include a focus group, which this student helped to plan and facilitate. This extra effort afforded a second chance for the student and provided further insight to the shelter. Stacy emphasizes the serious nature of evaluation to her students, noting that service learning is intended to benefit students, clients, and the agency. As they learn from the applied assignment, students are reminded that "this organization will make decisions based on these results and what you tell them." Clearly, Stacy sees program evaluation as rigorous and challenging, but she believes it also "can be a whole lot of fun," as it allows her to see many different organizations, connect with many different people, and have a chance to help.

Several points stand out as she talks about her classes. Along the lines of the case study, engaging stakeholders in productive ways is clearly one of her highest priorities. This includes both helping them make use of the evaluation and grounding the evaluation in an accurate understanding of the program. She notes that "one of the most dangerous things is when you don't know what you don't know," an especially relevant risk for external evaluators. But she also points out to her students that evaluators have the "luxury of time to consider many perspectives" and to discover patterns that may not be obvious to those too close to the program.

Source: S. Deck (personal communication, January 4, 2018).

SUMMARY AND POSSIBLE TRENDS FOR PROGRAM EVALUATION

Evaluators have become increasingly active in working with agencies and organizations whose programs have been evaluated. No longer do evaluators slip 10 copies of a completed evaluation into the mail after preparing a report in their offices. Instead, they seek to help the organization make use of an evaluation even when the findings are mixed or negative. The crucial issue is to encourage stakeholders to look upon an evaluation as a development tool to use in making program improvements. The participation of evaluators in planning discussions improves the chances for the appropriate use of feedback on needs, implementation, or outcomes.

The discipline of program evaluation has changed markedly from the 1960s, when it began its period of rapid growth. The idea of evaluating the effects of programs and policies is no longer novel or controversial in service agencies and government offices. Contemporary evaluators are far more sophisticated in methodological skills and in working with organizations than our disciplinary forbears were. And unlike evaluators in the not-very-distant past, we are more likely to expect confusion in program design and limitations in implementation in the field and to recognize the constraints on the activities of the managers of

service organizations and government agencies. We also see more clearly that each program setting is unique; what works in one place is not necessarily appropriate for another community or with another population. Evaluators have come to recognize the limitations of evaluation methods as well. The combination of a greater sensitivity to stakeholder needs, a deeper respect for the importance of the context of a program, and a less grandiose view of evaluation methods has resulted in better evaluations. Today, a smaller proportion of evaluations are inappropriately designed or superficially conducted than in the past. Nevertheless, being useful to organizations and the people served remains a challenge: we have our work cut out for us.

STUDY QUESTIONS

1 Students often seem puzzled by resistance to evaluation findings on the part of stakeholders responsible for providing a service. To develop an understanding of this resistance, imagine something you really worked at—making a high school team, writing an article for a school newspaper, or learning a piece of music, for example. Now imagine one of your parents criticizing you for not doing it perfectly without offering a word of praise. How would you feel? If you can put yourself into this situation, you might better understand the reaction of program personnel toward an insensitive evaluator.

2 If you are up to it, go back and imagine your hypothetical critical parent making a recommendation on how "you should have" reacted on the last football play, or developed the second paragraph differently, or smiled more when you played at your recital. Be sure to recognize that this hypothetical parent has never played football, written a newspaper article, or performed music him/herself. How likely are you to accept the recommendations? Now consider how a police officer or a mental health therapist would react to recommendations from a critical evaluator who has never been ordered to stop and question a possibly dangerous person or tried to counsel a depressed teenager threatening suicide.

3 What aspect of a person's demeanor would encourage you to consider recommendations made for helping you improve your work?

4 Suppose that you completed an evaluation of that sexual assault program mentioned in the first chapter. Suppose that you did make a good effort and learned (a) that both male and female participants were clearer on the definitions of what is and what is not a sexual assault; (b) that at the follow-up only a few female participants were taking the precautions recommended in the program; and (c) that at the follow-up most men indicated that they thought most sexual assaults were the victims' fault. What implications are suggested by these findings and how can you as the evaluator encourage utilization of your findings?

5 A college committee considered recommending approval of a proposed multidisciplinary program, say, psychology and law. What might the committee learn from a needs assessment? What might they use from a study of the jobs or further study of the alumni of the college's undergraduate programs? How could the committee encourage a learning climate among their colleagues in the college when the recommended proposal is presented to a larger, decision-making body?

6 Think of situations in which value polarization might occur. What might happen if a faculty and student committee considers whether grade inflation is a problem at its college?

ADDITIONAL RESOURCE

Gill, S. J. (2010). *Developing a learning culture in a nonprofit organization*. Thousand Oaks, CA: Sage Publications.

Methods are discussed that encourage the development of a culture of learning that uses reflection and feedback on both successes and failures. Gill shows how to use information from evaluations to contribute to a strong agency that motivates both staff and volunteers. In addition, the barriers that can stand in the way of a learning culture are discussed.

References

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Abry, T., Hulleman, C. S., & Rimm-Kaufman, S. (2015). Using indices of fidelity to intervention core components to identify program active ingredients. *American Journal of Evaluation, 36*(4), 320–338. doi:10.1177/1098214014557009

Ackoff, R. L. (1986). *Management in small doses*. New York, NY: Wiley.

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J. L., & Hsiung, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation. *Evaluation Review, 22*, 207–243.

Allen, C., Glasziou, P., & Del Mar, C. B. (1999). Bed rest: A potentially harmful treatment needing more careful evaluation. *Lancet, 354*, 1229–1233.

Altschuld, J. W. (Ed.). (2010). *The needs assessment kit*. Thousand Oaks, CA: Sage Publications.

Altschuld, J. W., & Engle, M. (2015). The inexorable historical press of the developing evaluation profession. In J. W. Altschuld & M. Engle (Eds.), *New Directions for Evaluation: No. 145. Accreditation, certification, and credentialing: Relevant concerns for U.S. evaluators*. San Francisco, CA: Jossey Bass.

Altschuld, J. W., & Watkins, R. (2014). A primer on needs assessment: More than 40 years of research and practice. In J. W. Altschuld & R. Watkins (Eds.), *New Directions for Evaluation: No. 144. Needs assessment: Trends and a view toward the future*. San Francisco, CA: Jossey-Bass.

Altschuld, J. W., & Witkin, B. R. (2000). *From needs assessment to action: Transforming needs into solution strategies*. Thousand Oaks, CA: Sage Publications.

American Evaluation Association (2017). Guiding principles for evaluators. *American Journal of Evaluation, 38*, 313–314.

American Evaluation Association. (2011). *The public statement on cultural competence in evaluation*. Retrieved from http://www.eval.org/p/cm/ld/fid=92.

American Evaluation Association. (2004). *Guiding principles for evaluators*. Retrieved from http://www.eval.org/Publications/GuidingPrinciplesPrintable.asp

American Psychological Association (APA). (2010). *Ethical principles of psychologists and code of conduct*. Retrieved from http://apa.org/ethics/code/index.aspx

American Educational Research Association. (2002). *Ethical standards of the American Educational Research Association*. Washington, DC: Author.

American Sociological Association. (1999). *Code of ethics*. Washington, DC: Author.

Anderson, C. A. (2001). Heat and violence. *Current Directions in Psychological Science, 10*, 33–38.

Angelo, T. A., & Cross, K. P. (1993). *Classroom assessment techniques: A handbook for college teachers* (Rev. ed.). San Francisco, CA: Jossey-Bass.

Argyris, C. (2004). *Reasons and rationalizations: The limits of organizational knowledge*. New York, NY: Oxford University Press.

Arvey, R. D., & Cole, D. A. (1989). Evaluating change due to training. In I. Goldstein & associates (Eds.), *Training and development in organizations*. San Francisco, CA: Jossey-Bass.

Aspelmeier, J., & Pierce, T. (2009). *SPSS: User friendly approach*. New York, NY: Worth Publishers.

Atkinson, C. M. (2009). *Jeremy Bentham: His life and work (1905)*. Charleston, SC: BiblioLife.

Averch, H. A. (1994). The systematic use of expert judgment. In J. S. Wholey, H. R. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation*. San Francisco, CA: Jossey-Bass.

Babbie, E. (2009). *Practice of social research* (12th ed.). Belmont, CA: Wadsworth.

Ball, S., & Bogartz, G. A. (1970). *The first year of Sesame Street*. Princeton, NJ: Educational Testing Service.

Bauer, S. M., & Toms, K. (1989). *Will there be any scientists in the class of 2000?* [Summary]. Proceedings of the Annual Meeting of the American Evaluation Association, 26.

Berk, R. A. (1977). Discretionary methodology decisions in applied research. *Sociological Methods and Research, 5*, 317–334.

Berk, R. A., & Rossi, P. H. (1976). Doing good or worse: Evaluation research politically reexamined. *Social Problems, 23*, 337–349.

Bickman, L. (2000). Summing up *program theory*. In P. J. Rogers, T. A. Hacsi, A. Petrosino, & T. A. Huebner (Eds.), *Program theory in evaluation: Challenges and opportunities. New directions for program evaluation* (Vol. 87, pp. 103–112). San Francisco, CA: Jossey-Bass.

Bingham, S. G., & Scherer, L. L. (2001). The unexpected effects of a sexual harassment educational program. *Journal of Applied Behavioral Science, 37*, 125–153.

Boruch, R. F. (1997). *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: Sage Publications.

Botcheva, L., White, C. R., & Huffman, L. C. (2003). Learning culture and outcomes measurement practices in community agencies. *American Journal of Evaluation, 22*, 421–434.

Bowen, M. L. & Tillman, A. S. (2015). Developing culturally responsive surveys. *American Journal of Evaluation, 36*(1), 25–41. doi:10.1177/1098214014539831

Brandon, P. R. & Fukunaga, L. L. (2014). The state of the empirical research literature on stakeholder involvement in program evaluation. *American Journal of Evaluation, 35*(1), 26–44. doi:10.1177/1098214013503699

Brinkerhoff, R. O. (2003). *The success case method*. San Francisco, CA: Berrett-Koehler.

Bronson, P., & Merryman, A. (2009). *NurtureShock: New thinking about children*. New York, NY: Twelve.

Brotman, B. (1983, January 2). "Workfare": What state terms success others call boondoggle. *Chicago Tribune, 3*, 1, 4.

Brown, J. B., & Adams, M. E. (1992). Patients as reliable reporters of medical care process. *Medical Care, 30*, 400–411.

Bryant, F. B., & Guilbault, R. L. (2002). "I knew it all along" eventually: The development of hindsight bias in reaction to the Clinton impeachment verdict. *Basic and Applied Social Psychology, 24*, 27–41.

Bryk, A. S., (Ed.). (1983). *Stakeholder-based evaluation*. San Francisco, CA: Jossey-Bass.

Bunda, M. A. (1983). Alternative ethics reflected in education and evaluation. *Evaluation News, 4*(1), 57–58.

Cagle, L. T., & Banks, S. M. (1986). The validity of assessing mental health needs with social indicators. *Evaluation and Program Planning, 9*, 127–142.

Caidin, M. (1960). *Let's go flying!* New York, NY: Dutton.

Campbell, D. T. (1983). *The problem of being scientific in program evaluation*. Paper presented at the meeting of the Evaluation Research Society. October, Chicago.

Campbell, D. T. (1987). Guidelines for monitoring the scientific competence of preventive intervention research centers. *Knowledge: Creation, Diffusion, Utilization, 8*, 389–430.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand-McNally.

Canadian Evaluation Society. (2010). Competencies for Canadian evaluation practice. Retrieved from http://www.evaluationcanada.ca/txt/2_competencies_cdn_evaluation_practice.pdf.

Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools: Now all states are above the national average*. Daniels, WV: Friends of Education.

Card, J. J., Greeno, C., & Peterson, J. L. (1992). Planning an evaluation and estimating its cost. *Evaluation & the Health Professions, 15*, 75–89.

Carey, R. G. (1974). Emotional adjustment in terminal patients. *Journal of Counseling Psychology, 21*, 433–439.

Carey, R. G. (1979). Evaluation of a primary nursing unit. *American Journal of Nursing, 79*, 1253–1255.

Carey, R. G. (2003). *Improving healthcare with control charts: Basic and advanced SPC methods and case studies*. Milwaukee, WI: ASQ Quality Press.

Carley, M. (1981). *Social measurement and social behaviors*. London: George Allen & Unwin.

Castner, L., & Cody, S. (1999). *Trends in FSP participation rates: Focus on September 1997*. Washington, DC: Mathematica Policy Research.

Caulkins, J. P., Rydell, C. P., Everingham, S. S., Chiesa, J., & Bushway, S. (1999). *An ounce of prevention, a pound of cure: The cost-effectiveness of school-based drug prevention programs*. Santa Monica, CA: Rand.

Champney, T. F. (1993). Comments on the use of multiple roles to maximize the utilization of outcome research. *Evaluation and Program Planning, 16*, 335, 336.

Chan, K-C. (1994). *Vaccines for children: Critical issues in design and implementation (T-PEMD-94-28)*. Washington, DC: U.S. General Accounting Office.

Chan, K-C. (1995). *Vaccines for children: Reexamination of program goals and implementation needed to insure vaccination* (T-PEMD-95-22). Washington, DC: U.S. General Accounting Office.

Chelimsky, E. (1997). The coming transformation in evaluation. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook*. Thousand Oaks, CA: Sage Publications.

Cholesterol screening. (1990). Washington, DC: Office of Inspector General, U.S. Department of Health and Human Services.

Chong, J., Gero, A., & Treichel, P. (2015). What indicates improved resilience to climate change? A learning and evaluative process developed from a child-centered, community based project in the Philippines. In D. Bours, C. McGinn, & P. Pringle (Eds.), *New Directions for Evaluation: No. 147. Monitoring and evaluation of climate change adaptation: A review of the landscape*. San Francisco, CA: Jossey-Bass.

Cicarelli, V. G., Cooper, W. H., & Granger, R. L. (1969). *The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development*, Westinghouse Learning Corporation, OEO Contract No, B89 4536.

Clark, C. (1989). Clear-cut economies: Should we harvest everything now? *The Sciences, 29*(1), 16–19.

Cohen, J. (1987). *Statistical power analysis for behavioral sciences* (Rev. ed.). Hillsdale, NJ: Lawrence Erlbaum.

Conner, R. F., Jacobi, M., Altman, D. G., & Aslanian, C. B. (1985). Measuring need and demand in evaluation research. *Evaluation Review, 9*, 717–734.

Cook, T. C. (2006). Describing what is special about the role of experiments in contemporary educational research: Putting the "Gold Standard" rhetoric into perspective. *Journal of Multidisciplinary Evaluation (online), No. 6*. Retrieved from http://survey.ate.wmich.edu/jmde/index.php/jmde_1/article/view/36/43.

Cook, T. D. (2000). The false choice between theory-based evaluation and experimentation. In P. J. Rogers, T. A. Hacsi, A. Petrosino, & T. A. Huebner (Eds.), *Program theory in evaluation: Challenges and opportunities. New directions for program evaluation* (No. 87). San Francisco, CA: Jossey-Bass.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation*. Chicago, IL: Rand-McNally.

Cook, T. D., & Devine, E. C. (1982). *Trying to discover explanatory processes through meta-analysis*. Paper presented at the National Meeting of the American Educational Research Association. March, New York.

Cook, T. D., & Shadish, W. R. (1994). Social experiments: Some developments over the past fifteen years. *Annual Review of Psychology, 45*, 545–580.

Cook, T. D., Appleton, H., Conner, R. F., Shaffer, A., Tamkin, G., & Webber, S. J. (1975). *Sesame Street revisited*. New York, NY: Russell Sage Publications.

Cook, T. D., Leviton, L. C., & Shadish, W. R. (1985). Program evaluation. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (3rd ed.). New York, NY: Random House.

Corbin, J. M., & Strauss, A. L. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Cordray, D. S. (1986). Quasi-experimental analysis: A mixture of methods and judgment. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design. New directions for program evaluation* (No. 31). San Francisco, CA: Jossey-Bass.

Crano, W. D., & Brewer, M. B. (1986). *Principles and methods of social research*. Newton, MA: Allyn and Bacon.

Crompton, D. W. (2009). Managing studies versus managing for evaluation capacity building. In D. W. Compton & M. Braizerman (Eds.), *Managing program evaluation: Towards explicating a professional practice. New directions for program evaluation* (Vol. 121, pp. 55–69). San Francisco, CA: Jossey-Bass.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*, 170–180.

Cummings, S. R., Rubin, S. M., & Oster, G. (1989). The cost-effectiveness of counseling smokers to quit. *JAMA, 261*, 75–79.

Datta, L. (1976). The impact of the Westinghouse/Ohio evaluation on the development of project Head Start. In C. C. Abt (Ed.), *The evaluation of social programs*. Beverly Hills, CA: Sage Publications.

Datta, L. (1994). Paradigm wars: A basis for peaceful coexistence and beyond. In C. S. Reichardt & S. F. Rallis (Eds.), *The qualitative-quantitative debate: New perspectives. New directions for program evaluation* (No. 61). San Francisco, CA: Sage Publications.

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth*. New York, NY: The Free Press.

Deck, S. M., & Thompson, T. (2014). *Metro United Way Financial Empowerment Pilot Program Evaluation Interim Report #1*, Louisville, KY: Author.

Defriese, G. B. (1990). Theory as method. In L. Sechrest, E. Perrin, & J. Bunker (Eds.), *Research methodology: Strengthening causal interpretations of nonexperimental data*. Rockville, MD: Department of Health and Human Services, Agency for Health Care Policy and Research, (PHS) 90-3454.

Deming, W. E. (1986). *Out of the crisis*. Cambridge, MA: MIT-CAES.

Dennis, M. L. (1994). Ethical and practical randomized field experiments. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation*. San Francisco, CA: Jossey Bass.

Denzin, N. K., & Lincoln, Y. S., eds. (2005). *Handbook of qualitative research* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Derby, A. (1989). Equating death and dollars on the highway. *Business and Society Review, 71*(*Fall*), 47, 48.

Devarajan, S., Dollar, D., & Holmgren, T. (2000). *Aid and reform in Africa: Lessons from ten case studies*. Washington, DC: The World Bank.

Dial, M. (1994). The misuse of evaluation in educational programs. In C. J. Stevens & M. Dial (Eds.), *Preventing the misuse of evaluation. New directions for program evaluation* (No. 64). San Francisco, CA: Jossey-Bass.

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

Dragon Naturally Speaking, 15 (2018). Burlington, MA: Nuance Communications.

Duncan, B. (2012). The partners for change outcome management system (PCOMS): The heart and soul of change project. *Canadian Psychology/Psychologie canadienne, 53*, 93–104. doi:10.1037/a0027762

Duncan, S. H., & Dickie, I. (2013) Family Group Conferencing: A pilot project within the Juvenile Court System in Louisville, Kentucky. *The Prevention Researcher, 20*, 11–14.

Durand, R., Decker, P. J., & Kirkman, D. M. (2014). Evaluation methodologies for estimating the likelihood of program implementation failure. *American Journal of Evaluation, 35*(3), 404–418. doi:10.1177/1098214014523824

Dyrbye, L. N., West, C. P., Satele, D., Boone, S., Tan, L., Sloan, J., & Shanafelt, T.D. (March 2014). Burnout among U.S. medical students, residents, and early career physicians relative to the general U.S. population. *Academic Medicine, 89*, 443–451. doi:10.1097/ACM.0000000000000134

Easterly, W. (2007). The ideology of development. *Foreign Policy, 161*, 30–35.

Eddy, D. M. (1990). Practice policies: Where do they come from? *JAMA, 263*, 1265, 1269, 1272, 1275.

Edmondson, A. C. (1996). Learning from mistakes is easier said than done: Group and organizational influences on the detection and correction of human error. *Journal of Applied Behavioral Science, 32*, 5–28.

Edmondson, A. C. (2004). Learning from mistakes is easier said than done. *Journal of Applied Behavioral*

Science, 40, 66–90.

Egan, G. (1988a). *Change-agent skills A: Assessing and designing excellence*. San Diego, CA: University Associates.

Egan, G. (1988b). *Change-agent skills B: Managing innovation and change*. San Diego, CA: University Associates.

Eisenberg, L. (1977). The social imperatives of medical research. *Science, 198*, 1105–1110.

Epley, N., & Dunning, D. (2000). Feeling "holier than thou": Are self-serving assessments produced by errors in self- or social prediction? *Journal of Personality and Social Psychology, 79*, 861–875.

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

Etzioni, A. (1994, July). Incorrigible. *Atlantic, 274*, 14–15.

Evans, R. G., & Robinson, G. C. (1980). Surgical day care: Measurements of economic payoff. *CMA Journal, 123*, 873–880.

Evans, R. I., & Raines, B. E. (1990). Applying a social psychological model across health promotion interventions. In J. D. Edwards, R. S. Tindale, L. Health, & E. J. Posavac (Eds.), *Social influence processes and prevention*. New York, NY: Plenum.

Fals-Stewart, W., Klostermann, K., & Yates T. (2005). Assessing the costs, benefits, cost-benefit ratio, and cost-effectiveness of marital and family treatments: Why we should and how we can. *Journal of Family Psychology, 19*, 28–39.

Faulkner, L., Ayers, J., & Huq, S. (2015). Meaningful measurement for community-based adaptation. In D. Bours, C. McGinn, & P. Pringle (Eds.), *New Directions for Evaluation: No. 147. Monitoring and evaluation of climate change adaptation: A review of the landscape*. San Francisco, CA: Jossey-Bass.

Felson, M. (1993). Social indicators for criminology. *Journal of Research in Crime and Delinquency, 30*, 400–411.

Felson, R. B. (1984). The effects of self-appraisal of ability on academic performance. *Journal of Personality and Social Psychology, 47*, 944–952.

Ferrari, J. R., Drexler, T., & Skarr, J. (2015). Finding a spiritual home: A pilot study on the effects of a spirituality retreat and loneliness among urban homeless adults. *Psychology, 6*, 210–216. doi:10.4236/psych.2015.63020.

Ferrari, J. R., Excell, S., Pardo, M. A., Crum, K. A., Dao, T. N., Dao, A., … Mortenson, E. (2017). A pilot assessment of a spiritual retreat program for homeless adults in recovery: Finding one's self in a safe, sober, and sacred place. *Journal of Addiction Research, 1*, 1–4.

Ferriss, A. L. (1988). Uses of social indicators. *Social Forces, 66*, 601–617.

Fetterman, D. M., & Wandersman, A., (Eds.). (2004). *Empowerment evaluation principles in practice*. New York, NY: The Guilford Press.

Figueredo, A. J. (1993). Critical multiplism, meta-analysis, and generalization: An integrative commentary. In L. Sechrest (Ed.), *Program evaluation: A pluralistic enterprise. New directions for program evaluation* (No. 60). San Francisco, CA: Jossey-Bass.

Fine, M., & Vanderslice, V. (1992). Qualitative activist research: Reflections on methods and politics. In F. B. Bryant, J. D. Edwards, R. S. Tindale, E. J. Posavac, L. Heath, Y. Suarez-Balcazar, & E. Henderson (Eds.), *Methodology in applied social psychology*. New York, NY: Plenum.

Fitch, R. W. (1995). Accepting the null hypothesis. *Memory and Cognition, 23*, 132–138.

Fitzpatrick, J. (2002). Dialogue with Stewart Donaldson. *American Journal of Evaluation, 23*, 347–365.

Fleischer, D. N., & Christie, C. A. (2009). Results from a survey of U.S. American Evaluation Association members. *American Journal of Evaluation, 30*, 158–175.

Fowler, F. J., Jr., & Cosenza, C. (2009). Design and evaluation of survey questions. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (Rev. ed.). Thousand Oaks, CA: Sage Publications.

Freiman, J. A., Chalmers, T. C., Smith, H., Jr., & Kuebler, R. R. (1978). The importance of beta, the type II

error and sample size in the design and interpretation of the randomized control trial. *New England Journal of Medicine, 299*, 690–694.

French, M. T. (2000). Economic evaluation of alcohol treatment services. *Evaluation and Program Planning, 23*, 27–39.

Fross, K., Cracknell, B., & Samset, K. (1994). Can evaluation help an organization to learn? *Evaluation Review, 18*, 574–591.

Fuller, R. K., Branchey, L., Brightwall, D. R., Derman, R. M., Emrick, C. D., Iber, F. L., … Shaw, S. (1986). Disulfiram treatment for alcoholism: A Veterans Administration cooperative study. *JAMA, 68*, 1449–1455.

Gabielkov, M., Ramachandran, A., Chaintreau, A., & Legout, A. (2016). Social clicks. *ACM SIGMETRICS Performance Evaluation Review (ACM Digital Library), 44*, 179–192.

Galport, N., & Azzam, T. (2017). Evaluator training needs and competencies. *American Journal of Evaluation, 38*(1), 280–100. doi:10.1177/1098214016643183

Garland, A. F., Haine, R. A., & Boxmeyer, C. L. (2007). Determinates of youth and parent satisfaction in usual care psychotherapy. *Evaluation and Program Planning, 30*, 45–54.

Gates, E., & Dyson, L. (2017). *Implications of the changing conversation about causality for evaluators, American Journal of Evaluation, 38*(1), 29–46. doi:10.1177/1098214016644068

Gehrs, M., Smith-Fowler, H., Rourke, S. B., Wasylenki, D., Smith, M., & Cousins, J. B. (2004). Inside the black box: Challenges in implementation evaluation of community mental health case management programs. *Canadian Journal of Program Evaluation, 19* (*Special issue*), 109–133.

Gerstein, D. R., Johnson, R. A., Harwood, H. J., Fountain, D., Suter, N., & Malloy, K. (1994). *Evaluating recovery services: The California drug and alcohol treatment (CALDATA)*. Sacramento, CA: California Department of Alcohol and Drug Programs.

Gilbert, J. P., Light, R. J., & Mosteller, F. (1975). Assessing social innovations: An empirical base for policy. In A. R. Lumsdaine & C. A. Bennett (Eds.), *Evaluation and experiment*. New York, NY: Academic Press.

Gill, S. J. (2010). *Developing a learning culture in a nonprofit organization*. Thousand Oaks, CA: Sage Publications.

Gillmore, G. M., & Greenwald, A. G. (1999). Using statistical adjustment to reduce the biases in student ratings. *American Psychologist, 54*, 518–519.

Glasgow, R. E., Terborg, J. R., Strycker, L. A., Boles, S. M., & Hollis, J. F. (1997). Take Heart II: Replication of a worksite health promotion trial. *Journal of Behavioral Medicine, 20*, 143–161.

Glenmullen, J. (2000). *Prozac backlash*. New York, NY: Simon & Schuster.

Goering, P. N., & Wasylenki, D. A. (1993). Promoting the utilization of outcome study results by assuming multiple roles within an organization. *Evaluation and Program Planning, 16*, 329–334.

Goldman, M., & Beaumont, T. (2017). A real world evaluation of a treatment for infant colic based on the experience and perceptions of 4004 parents. *Journal of Health Visiting, 5*, 134–142.

"Good News—crime is up!" (1983, May 8). *Chicago Tribune*, Sec. 2, p. 2.

Green, J., & Wintfeld, N. (1993). How accurate are hospital discharge data for evaluating effectiveness of care? *Medical Care, 31*, 719–731.

Green, R. S., & Jerrell, J. M. (1994). The generalizability of brief ratings of psychosocial functioning. *Evaluation and Program Planning, 17*, 141–151.

Greene, J. C. (1987). Stakeholder participation in evaluation: Is it worth the effort? *Evaluation and Program Planning, 10*, 379–394.

Greiner, J. M. (1994). Use of ratings by trained observers. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation*. San Francisco, CA: Jossey-Bass.

Gross, D. M., & Scott, S. (1990, July 16). Proceeding with caution. *Time, 56*–62.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage Publications.

Guilbault, R. L., Bryant, F. B., Brockway, J. H., & Posavac, E. J. (2004). A meta-analysis of research on hind-sight bias. *Basic and Applied Social Psychology, 26*, 103–117.

Hahn, C. K., Morris, J. M., & Jacobs, G. A. (2017). Predictors of bystander behaviors and sexual assertiveness among college women attending a sexual assault prevention program. *Journal of Community Psychology, 45*, 672–677.

Harmsen, I. A., Mollema, L., Ruiter, R. C., Paulussen, T. W., de Melker, H. E., & Kok, G. (2013). Why parents refuse childhood vaccination: A qualitative study using online focus groups. *BMC Public Health, 131183*. doi:10.1186/1471-2458-13-1183

Hartog, J. (1999). Over- and under-education and the relation to vocational training. *Vocational Training: European Journal, 16*, 47–52.

Hatry, H. P. (1994). Collecting data from agency records. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation*. San Francisco, CA: Jossey-Bass.

Hatry, H., Wholey, J. S., & Newcomer, K. E. (2004). Other issues and trends in evaluation. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (Rev. ed.). San Francisco, CA: Jossey Bass.

Heath, C., & Heath, D. (2008). *Made to stick: Why some ideas survive and others die*. New York, NY: Random House.

Heath, L., & Petraitis, J. (1986). Television viewing and fear of crime: Where is the scary world? *Basic and Applied Social Psychology, 8*, 97–123.

Hedrick, T. E. (1994). The quantitative-qualitative debate: Possibilities for integration. In C. S. Reichardt & S. F. Rallis (Eds.), *The qualitative-quantitative debate: New perspectives. New directions for program evaluation* (No. 61). San Francisco, CA: Jossey-Bass.

Hegarty, T. W., & Sporn, D. L. (1988). Effective engagement of decisionmakers in program evaluation. *Evaluation and Program Planning, 11*, 335–340.

Henderson, G. R., Mead, G. E., van Dijke, M. L., Ramsay, S., McDowall, M. A., & Dennis, M. (2008). Use of statistical process control charts in stroke medicine to determine if clinical evidence and changes in service delivery were associated with improvements in the quality of care. *Quality and Safety in Health Care, 17*, 301–306.

Henderson, M., Johnson, N. F., & Auld, G. (2013). *Silences of ethical practice: Dilemmas for researchers using social media, Educational Research and Evaluation, 19*, 546–560.

Hendricks, M. (1994). Making a splash: Reporting evaluation results effectively. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation*. San Francisco, CA: Jossey Bass.

Herman, P. M., Avery, D. J., Schemp, C. S., & Walsh, M. E. (2009). Are cost-inclusive evaluations worth the effort? *Evaluation and Program Planning, 32*, 55–61.

"High-rise brought low at last." (1998, July 11). *The Economist, 31*–32.

Hilkevitch, J. (2000, January 15). Whistle-blower plan for air safety. *Chicago Tribune*, Sec. 1, p. 2.

Hogan, R. R. (1985). Response bias in a student follow-up: A comparison of low and high return surveys. *College and Universities, 61*, 17–25.

Horton, S. V. (1987). Reduction of disruptive mealtime behavior by facial screening. *Behavior Modification, 11*, 53–64.

Hougland, J. G., Jr. (1987). Criteria for client evaluation of public programs: A comparison of objective and perceptual measures. *Social Science Quarterly, 68*, 386–394.

House, P. W., & Shull, R. D. (1988). *Rush to policy: Using analytic techniques in public sector decision making*. New Brunswick, NJ: Transaction Books.

Howard, P. K. (1994). *The death of common sense: How law is suffocating America*. New York, NY: Random House.

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA, 294*(2), 218–228.

Jacobsen, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12–19.

James Bell Associates (2009). *Florida department of children and families—Waiver information packet*.

Arlington, VA: Author. Retrieved from http://www.jbassoc.com/reports/documents/fl%20handout-final.pdf.

Jason, L. A., & Liotta, R. F. (1982). Assessing community responsiveness in a metropolitan area. *Evaluation Review, 6*, 703–712.

Joglekar, P. N. (1984). Cost-benefit studies of health care programs: Choosing methods for desired results. *Evaluations & the Health Professions, 7*, 285–303.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards: How to assess evaluations of educational programs* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Judd, C. M., Kenny, D. A., & McClelland, G. H. (2008). Estimating and testing mediation and moderation in within participant designs. *Psychological Methods, 6*, 115–134.

Julnes, G., & Mohr, L. B. (1989). Analysis of no-difference findings in evaluation research. *Evaluation Review, 13*, 628–655.

Kahn, J. (2000, May 1). Is Harvard worth it? *Fortune, 141*, 200–204.

Kane, R. L., & Kane, R. A. (1978). Care of the aged: Old problems in need of new solutions. *Science, 200*, 913–919.

Karwath, R. (1990, August 28). Jury's out on pregnant teens plan. *Chicago Tribune*, Sec. 2, p. 4.

Katz, M. M., & Warren, W. L. (1998). *Katz Adjustment Scales Relative Report Form (KAS–R)*. Los Angeles, CA: Western Psychological Services.

Kent, D. M., Fendrick, A. M., & Langa, K. M. (2004). New and dis-improved: On the evaluation and use of less effective, less expensive medical interventions. *Medical Decision Making, 24*, 282–286.

Kibel, B. M. (1999). *Success stories as hard data: An introduction to results mapping*. New York, NY: Kluwer Academic.

King, J. (2017). Using economic methods evaluatively. *American Journal of Evaluation, 38*, 101–113. doi:10.1177/1098214016641211

King, J. A. (1994). Meeting the educational needs of at-risk students: A cost analysis of three models. *Educational Evaluation and Policy Analysis, 16*, 1–19.

King, J. A., & Stevahn, L. (2013). *Interactive evaluation practice: Mastering the interpersonal dynamics of program evaluation*. Thousand Oaks, CA: Sage Publications.

King, J. A., & Stevahn, L. (2015). Competencies for program evaluators in light of adaptive action: What? So what? Now what? In J. W. Altschuld & M. Engle (Eds.), *New Directions for Evaluation: No. 145. Accreditation, certification, and credentialing: Relevant concerns for U.S. evaluators*. San Francisco, CA: Jossey-Bass.

King, J. A., Stevahn, L., Ghere, G., & Minnema, J. (2001). Toward a taxonomy of essential evaluator competencies. *American Journal of Evaluation, 22*, 229–247.

Kingsbury, K. (2008). The value of a human life. *Time* (online), Retrieved from http://www.time.com/time/health/article/0,8599,1808049,00.html.

Klerman, J. A., Olsho, L. E. W., & Bartlett, S. (2015). Regression discontinuity in prospective evaluations. *American Journal of Evaluation, 36*, 403–416. doi:10.1177/1098214014553786

Klitzner, M., Gruenewald, P. J., Bamberger, E., & Rossiter, C. (1994). A quasi-experimental evaluation of Students Against Drunk Driving. *American Journal of Drug and Alcohol Abuse, 20*, 57–74.

Koelsch, L. D., Brown, A. L., & Boisen, L. (2016). Bystander perceptions: Implications for university sexual assault prevention programs. In R. D. Maiuro (Ed), *Perspectives on college sexual assault: Perpetrator, victim, and bystander* (pp. 197–211). New York, NY: Springer.

Kohfeldt, D., Bowen, A. R., & Langhout, R. D. (2016). 'They think kids are stupid': yPAR and confrontations with institutionalized power as contexts for children's identity work. *Revista Puertorriqueña De Psicología, 27*, 276–291.

Kramer, P. D. (1993). *Listening to Prozac*. New York, NY: Viking.

Krueger, R. A. (1994). *Focus Groups: A Practical Guide for Applied Research*, (2nd Ed.) Thousand Oaks, CA: Sage.

Krueger, R. A., & Casey, M. A. (2000). *Focus groups: A practical guide for applied research* (3rd ed.). Thousand

Oaks, CA: Sage Publications.

Kuji-Shikatani, K. (2015). Credentialed evaluator designation program, the Canadian experience. In J. W. Altschuld & M. Engle (Eds.), *New Directions for Evaluation, 145, Accreditation, certification, and credentialing: Relevant concerns for U.S. evaluators*. San Francisco, CA: Jossey-Bass.

Kuo, E., Vander Stoep, A., McCauley, E., & Kernic, M. A. (2009). Cost-effectiveness of a school-based emotional health screening program. *Journal of School Health, 79*, 277–285.

Langhout, R. D. (2015). Considering community psychology competencies: A love letter to budding scholar-activists who wonder if they have what it takes. *American Journal of Community Psychology, 55*, 266–278. doi:10.1007/s10464-015-9711-5

Langhout, R. D., & Fernandez, J. S. (2015). Empowerment evaluation conducted by fourth- and fifth-grade students. In D. M. Fetterman, S. J. Kaftarian, & A. Wandersman (Eds.), *Empowerment evaluation: Knowledge and tools for self-assessment, evaluation capacity building, and accountability* (2nd ed.). Thousand Oaks, CA: Sage Publications.

LaVelle, J. M., & Donaldson, S. I. (2015). The state of preparing evaluators. In J. W. Altschuld & M. Engle (Eds.), *New Directions for Evaluation, 145, Accreditation, certification, and credentialing: Relevant concerns for U.S. evaluators*. San Francisco, CA: Jossey-Bass.

Lavrakas, P. J. (1998). Methods for sampling and interviewing for telephone surveys. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods*. Thousand Oaks, CA: Sage Publications.

Lazar, I. (1981, January). Early intervention is effective. *Educational Leadership, 303*–305.

Le Menestrel, S. M., Walahoski, J. S., & Mielke, M. B. (2014). A partnership model for evaluation. *American Journal of Evaluation, 35*, 61–72. doi:10.1177/1098214013506600

Leeuw, F. L. (2003). Reconstructing program theories: Methods available and problems to be solved. *American Journal of Evaluation, 24*, 5–20.

Leik, R. K., & Chalkley, M. A. (1990, May–June). Parent involvement: What is it that works? *Children Today, 34*–37.

Lepicki, T., & Boggs, A. (2014). Needs assessments to determine training requirements. In J. W. Altschuld & R. Watkins (Eds.), *New Directions for Evaluation, 144, Needs assessment: Trends and a view toward the future*. San Francisco, CA: Jossey-Bass.

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis* (Rev. ed.). Thousand Oaks, CA: Sage Publications.

Leviton, L. C. (2014). Generative insights from the Eleanor Chelimsky forum on evaluation theory and practice. *American Journal of Evaluation, 35*, 244–249. doi:10.1177/1098214013503701

Leviton, L. C., & Boruch, R. F. (1983). Contributions of evaluation in education programs and policy. *Evaluation Review, 7*, 563–598.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Lincoln, K. D. (2000). Social support, negative social interactions, and psychological well-being. *Social Service Review, 74*, 231–252.

Lincoln, Y. S. (1990). The making of a constructivist: A remembrance of transformations past. In E. G. Guba (Ed.), *The paradigm dialog*. Newbury Park, CA: Sage Publications.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage Publications.

Linfield, K., Wohlfarth, D., Gibson, M., & Jones, L. (2013). *Bridges to Tomorrow: Final Program Evaluation Report 2007–2013*. Louisville, KY: Author.

Linn, R. L. (2000, March). Assessments and accountability. *Educational Researcher, 29*, 4–16.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage Publications.

Lipsey, M. W. (1993). Theory as method: Small theories of treatment. In L. B. Sechrest & A. G. Scott (Eds.), *Understanding causes and generalizing about them. New directions in program evaluation* (No. 57). San Francisco, CA: Jossey-Bass.

Lipsey, M. W., & Cordray, D. S. (2000). Evaluation methods for social intervention. *Annual Review of*

Psychology, 51, 345–375.

Lipsey, M. W., Crosse, S., Dunkle, J., Pollard, J. A., & Stobart, G. (1985). Evaluation: The state of the art and the sorry state of the science. In D. S. Cordray (Ed.), *Utilizing prior research in evaluation planning. New directions for program evaluation* (No. 27). San Francisco, CA: Jossey-Bass.

Lipsey, M. W., Petrie, C., Weisburd, D., & Gottfredson, D. (2006). Improving evaluation of anti-crime programs. *Journal of Experimental Criminology, 2*, 271–307.

Lipsey, M. W., & Pollard, J. A. (1989). Driving toward theory in program evaluation: More models to choose from. *Evaluation and Program Planning, 12*, 317–328.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181–1209.

Louie, J., Rhoads, C., & Mark, J. (2016). Challenges to using the regression discontinuity design in educational evaluations. *American Journal of Evaluation, 37*, 381–407. doi:10.1177/1098214015621787

Love, A. J. (1986, October). Using evaluation to identify service gaps in mental health services to youth. Paper presented at the meeting of the American Evaluation Association. Kansas City, MO.

Lurigio, A. J., & Swartz, J. (1994). Life at the interface: Issues in the implementation and evaluation of a multiphased, multiagency jail-based treatment program. *Evaluation and Program Planning, 17*, 205–216.

MacRae, D., Jr. (2009). *Policy indicators: Links between social science and public debate*. Chapel Hill, NC: University of North Carolina Press Enduring Editions.

Mager, R. F. (1997). *Goal analysis* (3rd ed.). Atlanta, GA: CEP Press.

Malcolm, M. T., Madden, J. S., & Williams, A. E. (1974). Disulfiram implantation critically evaluated. *British Journal of Psychiatry, 125*, 485–489.

Malhotra, N. K. (2010). *Marketing research: An applied orientation* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Mangione, T. W. (1998). Mail surveys. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods*. Thousand Oaks, CA: Sage Publications.

Manoff, R. K. (1985). *Social marketing: New imperative for public health*. New York, NY: Praeger.

Margasak, L. (2001, June 6). Pentagon's anti-fraud auditors faked data. *Chicago Tribune*, Sec. 1, p. 9.

Mark, M. M., & Shotland, R. L. (Eds.). (1987). *Multiple methods in program evaluation*. San Francisco, CA: Jossey-Bass.

Marsh, H. W. (1998). Simulation study of nonequivalent group-matching and regression-discontinuity designs: Evaluations of gifted and talented programs. *Journal of Experimental Education, 66*, 163–192.

Mashaal, M. (2006). *Bourbaki: A secret society of mathematicians*. Providence, RI: American Mathematical Society.

Massey, O. T., & Wu, L. (1994). Three critical views of functioning: Comparisons of assessments made by individuals with mental illness, their case managers, and family members. *Evaluation and Program Planning, 17*, 1–7.

Mathison, S. (1994). Rethinking the evaluator role: Partnerships between organizations and evaluators. *Evaluation and Program Planning, 17*, 299–304.

Maxwell, G. S. (1985). Problems of being responsive: Reflections on an evaluation of a program for training motorcycle riders. *Evaluation and Program Planning, 8*, 339–348.

McClary, S., Lubin, B., Evans, C., Watt, B., & Lebedun, M. (1989). Evaluation of a community treatment program for young adult schizophrenics. *Journal of Clinical Psychology, 45*, 806–808.

McClave, T., Benson, P. G., & Sincich, T. (2008). *Statistics for business and economics* (10th ed.). Upper Saddle River, NJ: Prentice Hall.

McDavid, J. C., & Huse, I. (2015). How does accreditation fit into the picture? In J. W. Altschuld & M. Engle (Eds.), *New Directions for Evaluation: No. 145, Accreditation, certification, and credentialing: Relevant concerns for U.S. evaluators*. San Francisco, CA: Jossey-Bass.

McGarrell, E. F., & Sabath, M. J. (1994). Stakeholder conflict in an alternative sentencing program: Implications for evaluation and implementation. *Evaluation and Program Planning, 17, 179*–186.

McKillip, J. (1991). Effect of mandatory premarital HIV testing on marriage: The case of Illinois. *American*

Journal of Public Health, 81, 650–653.

McKillip, J. (1992). Research without control groups: A control construct design. In F. B. Bryant, J. Edwards, R. S. Tindale, E. J. Posavac, L. Heath, E. Henderson, & Y. Suarez-Balcazar (Eds.), *Methodological issues in applied social psychology*. New York, NY: Plenum Press.

McKillip, J. (1998). Need analysis: Process and techniques. In L. Bickman & D. J. Rog (Ed.), *Handbook of applied social research methods*. Thousand Oaks, CA: Sage Publications.

McKillip, J., & Baldwin, K. (1990). Evaluation of an STD education media campaign: A control construct design. *Evaluation Review, 14*, 331–346.

McKillip, J., Lockhart, D. C., Eckert, P. S., & Phillips, J. (1985). Evaluation of a responsible alcohol use media campaign on a college campus. *Journal of Alcohol and Drug Education, 30*(3), 88–97.

McKillip, J., Moirs, K., & Cervenka, C. (1992). Asking open-ended consumer questions to aid program planning: Variations in question format and length. *Evaluation and Program Planning, 15*, 1–6.

McKnight, J. (1995). *The careless society: Community and its counterfeits*. New York, NY: Basic Books.

McPhail, A., Pikula, H., Roberts, J., Browne, G., & Harper, D. (1990). Primary nursing: A randomized crossover trial. *Western Journal of Nursing Research, 12*, 188–200.

McSweeny, A. J., & Creer, T. L. (1995). Health-related quality-of-life in medical care. *Disease-a-Month, 41*, 1–71.

McWhorter, J. H. (2000). *Losing the race: Self-sabotage in black America*. New York, NY: Free Press.

Mead, L. M. (2016). On the "how" of research to get inside the black box. In L. R. Peck (Ed.), *New Directions for Evaluation: No. 152, Social experiments in practice: The what, why, when, where, and how of experimental design & analysis*. San Francisco, CA: Jossey-Bass.

"Medicare proposed payment system for end-stage renal disease would reduce spending, improve quality." (2009, September 15). Washington, DC: Centers for Medicare and Medicaid Services.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*, 108–141.

Meier, S. T. (2004). Improving design sensitivity through intervention-sensitive measures. *American Journal of Evaluation, 25*, 321–334.

Meier, S. T. (2008). *Measuring change in counseling and psychotherapy*. New York, NY: Guilford Press.

Mendez, D., & Warner, K. E. (2000). Smoking prevalence in 2010: Why the healthy people goal is unattainable. *American Journal of Public Health, 90*, 401–403.

Mertens, D. M., & Wilson, A. T. (2012). *Program evaluation theory and practice*. New York, NY: Guilford.

Miguel, E. M., Chou, T., Golik, A., Cornacchio, D., Sanchez, A. L., DeSerisy, M., & Comer, J. S. (2017). Examining the scope and patterns of deliberate self-injurious cutting content in popular social media. *Depression and Anxiety, 34*, 786–793.

Millenson, M. L. (1987, June 15). System puts doctors, cost cutters at odds. *Chicago Tribune*, Sec. 1, pp. 1, 11.

Millsap, R. E., & Hartog, S. B. (1988). Alpha, beta, and gamma change in evaluation research: A structural equation approach. *Journal of Applied Psychology, 73*, 574–584.

Mitra, A. (1994). Use of focus groups in the design of recreation needs assessment questionnaires. *Evaluation and Program Planning, 17*, 133–140.

Mitra, A. (1998). *Fundamentals of quality control and improvement* (Rev. ed.). Upper Saddle River, NJ: Prentice Hall.

Moore, T. J. (1995). *Deadly medicine*. New York, NY: Simon & Schuster.

Morgan, D. L., Eliot, S., Lowe, R. A., & Gorman, P. (2016). Dyadic interviews as a tool for qualitative evaluation. *American Journal of Evaluation, 37*, 109–117. doi:10.1177/1098214015611244

Morgan, M. G., Fischhoff, B., Bostrom, A., Lave, L., & Atman, C. J. (1992). Communicating risk to the public. *Environmental Science and Technology, 26*, 2048–2056.

Morris, M. (2015). More than the Beatles. *American Journal of Evaluation, 36*, 99–107. doi:10.1177/1098214014557808

Mosher, C. J., & Akins, S. A. (2007). *Drugs and drug policy: The control of consciousness alteration*. Thousand Oaks, CA: Sage Publications.

Moskowitz, J. M. (1993). Why reports of outcome evaluations are often biased or uninterpretable: Examples from evaluations of drug abuse prevention programs. *Evaluation and Program Planning, 16*, 1–10.

Mosteller, F. (1981). Innovation and evaluation. *Science, 211*, 881–886.

Mowbray, C. T., Bybee, D., Holter, M., & Lewandowski, L. (2006). Validation of a fidelity rating instrument for consumer-operated services. *American Journal of Evaluation, 27*, 9–27.

Muennig, P., & Bounthavong, M. (2016). *Cost effectiveness analysis in health: A practical approach* (3rd ed.). San Francisco, CA: Jossey-Bass.

Munter, C., Cobb, P., & Shekell, C. (2016). The role of program theory in evaluation research. *American Journal of Evaluation, 37*, 26–37. doi:10.1177/1098214015571122

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Nagel, S. S. (1983). Nonmonetary variables in benefit-cost evaluation. *Evaluation Review, 7*, 37–64.

Namey, E., Guest, G., McKenna, K., & Chen, M. (2016). Evaluating bang for the buck. *American Journal of Evaluation, 37*, 425–440. doi:10.1177/1098214016630406

National Center on Elder Abuse: Fact Sheet. (2005). Retrieved from http://www.ncea.aoa.gov/ncearoot/Main_Site/pdf/publication/FinalStatistics050331.pdf.

National Safety Council. (2009). *Report on injuries in America*. Retrieved from http://www.nsc.org/resources/library/report_table_1.aspx.

Neal, J. W., Neal, Z. P., VanDyke, E., & Kornbluh, M. (2015). Expediting the analysis of qualitative data in evaluation. *American Journal of Evaluation, 36*, 118–132. doi:10.1177/1098214014536601

Neuhauser, D. (1991). Parallel providers, ongoing randomization, and continuous improvement. *Medical Care, 29*(Suppl. 7), 5–8.

Newman, D. L., & Brown, R. D. (1996). *Applied ethics for program evaluation*. Thousand Oaks, CA: Sage Publications.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241–301.

Notebook. (1994, February 2). *Chronicle of Higher Education*, p. A31.

Office of Inspector General. (2009). *Reports: Office of evaluation and inspection*. Retrieved from https://oig.hhs/gov/oci/subject_index.asp.

On a diet? Don't trust your memory. (1989, October). *Psychology Today*, p. 12.

Overstreet, E. J., Grasso, A. J., & Epstein, I. (1993). Management information systems and external policy advocacy: The Boysville length of stay study. In A. J. Grasso & I. Epstein (Eds.), *Information systems in child, youth, and family agencies: Planning, implementation, and service enhancement*. New York, NY: Haworth Press.

Patton, M. Q. (1980). *Qualitative evaluation methods*. Beverly Hills, CA: Sage Publications.

Patton, M. Q. (2003). Qualitative evaluation checklist. *Evaluation checklists project*. Retrieved from https://wmich.edu/sites/default/files/attachments/u350/2014/qualitativeevalchecklist.pdf.

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

Patton, M. Q. (2014). What brain sciences reveal about integrating theory and practice. *American Journal of Evaluation, 35*, 237–244. doi:10.1177/1098214013503700

Patton, M. Q. (2016). What is essential in developmental evaluation? On integrity, fidelity, adultery, abstinence, impotence, long-term commitment, integrity, and sensitivity in implementing evaluation models. *American Journal of Evaluation, 37*, 250–265. doi:10.1177/1098214015626295

Paul, G. L. (Ed.). (1986). *Assessment in residential settings: Principles and methods to support cost-effective quality operations*. Champaign, IL: Research Press.

Paulos, J. A. (1988). *Innumeracy: Mathematical illiteracy and its consequences*. New York, NY: Hill and Wang.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. Thousand Oaks, CA: Sage Publications.

Pearce, D., & Markandya, A. (1988, April/May). Pricing the environment. *The OECD Observer*, pp. 23–26.

Peck, L. R. (2016). On the "how" of social experiments: Analytic strategies for getting inside the black box. In L. R. Peck (Ed.), *New Directions for Evaluation: No. 152. Social experiments in practice: The what, why,*

when, where, and how of experimental design & analysis. San Francisco, CA: Jossey-Bass.

Perkins, O. (2008, December 1). PolicyBridge ad campaign promotes education. *The Plan Dealer*, p. B1.

Peterson, R. D. (1986). The anatomy of cost-effectiveness analysis. *Evaluation Review, 10*, 29–44.

Pion, G. M., Cordray, D. S., & Anderson, S. (1993). Drawing the line between conjecture and evidence about the use and benefit of "practice" methodologies. *Professional Psychology: Research and Practice, 24*, 245–249.

Posavac, E. J. (1975). *A turning point: Survey of past residents in the clinical pastoral education program*. Park Ridge, IL: Lutheran General Hospital.

Posavac, E. J. (1992). Communication of applied social psychology: An art and a challenge. In F. B. Bryant, J. D. Edwards, R. S. Tindale, E. J. Posavac, L. Heath, Y. Suarez-Balcazar, & E. Henderson (Eds.), *Methodology in applied social psychology*. New York, NY: Plenum.

Posavac, E. J. (1995). Program quality and program effectiveness: A review of evaluations of programs to reduce excessive medical diagnostic testing. *Program Planning and Evaluation, 18*, 1–11.

Posavac, E. J. (1998). Toward more informative uses of statistics: Alternatives for program evaluators. *Evaluation and Program Planning, 21*, 243–254.

Posavac, E. J., & Carey, R. G. (2007). *Program evaluation: Methods and case studies* (7th ed.). Upper Saddle River, NJ: Pearson.

Posavac, E. J., & Hartung, B. M. (1977). An exploration into the reasons people choose a pastoral counselor instead of another type of psychotherapist. *The Journal of Pastoral Care, 31*, 23–31.

Posavac, E. J., & Sinacore, J. M. (1984). Reporting effect size in order to improve the understanding of statistical significance. *Knowledge: Creation, Diffusion, Utilization, 5*, 503–508.

Posavac, E. J., Sinacore, J. M., Brotherton, S. E., Helford, M., & Turpin, R. S. (1985). Increasing compliance to medical treatment regimens. *Evaluation & the Health Professions, 8*, 7–22.

Posavac, S. S. (1998). Strategic overbidding in contingent valuation: Stated economic value of public goods varies according to consumers' expectations of funding source. *Journal of Economic Psychology, 19*, 205–214.

Posavac, S. S., Sanbonmatsu, D. M., & Ho, E. A. (2002). The effects of subjective considerations of alternatives on consumer choice and attitude-decision consistency. *Journal of Consumer Psychology, 12*, 203–213.

Postal service may dump billion dollar parcel plan. (1990, June 14). *Chicago Tribune*, Sec. 4, p. 1.

Pouw, N., Dietz, T., Bélemvire, A., de Groot, D., Millar, D., Obeng, F., … Zaal, F. (2017). Participatory assessment of development interventions. *American Journal of Evaluation, 38*, 47–59. doi:10.1177/1098214016641210

Preskill, H. (1994). Evaluation's role in enhancing organization learning. *Evaluation and Program Planning, 17*, 291–297.

Preskill, H., & Coghlan, A. T. (Eds.). (2003). Using appreciative inquiry in evaluation. *New directions in program evaluation* (No. 100). San Francisco, CA: Jossey-Bass.

Press Ganey Associates. (2009). *Medical practice pulse report: Patient perspectives on American health care*. South Bend, IN: Author.

Price, A., Liew, S. M., Kirkpatrick, J., Price, J., Lopreto, T., & Nelken, Y. (2017). Mind the gap in clinical trials: A participatory action analysis with citizen collaborators, *Journal of Evaluation in Clinical Practice, 23*, 178–184.

Prochaska, J. O., & Norcross, J. C. (2001). Stages of change. *Psychotherapy: Theory, Research, Practice, Training, 38*, 443–448. doi:10.1037/0033-3204.38.4.443

Pulford, J., Adams, P., & Sheridan, J. (2008). Therapist attitudes and beliefs relevant to client dropout revisited. *Community Mental Health Journal, 44*, 181–186.

Raines, J. C. (2008). *Evidence based practice in school mental health*. New York, NY: Oxford University Press.

Rajkumar, A. S., & French, M. T. (1997). Drug abuse, crime costs, and the economic benefits of treatment. *Journal of Quantitative Criminology, 13*, 291–323.

Rawls, J. (2000). *A theory of justice* (Rev. ed.). Cambridge, MA: Harvard University Press.

Ray, M. (2006). Choosing a truly external evaluator. *American Journal of Evaluation, 27*, 372–377.

Raykov, T. (1999). Are simple change scores obsolete? An approach to studying correlates and predictors of change. *Applied Psychological Measurement, 23*, 120–126.

REACH Evaluation. (2014). *Open arms*. Louisville, KY: Author

Reichardt, C. S. (1979). The statistical analysis of data from nonequivalent group designs. In T. D. Cook & D. T. Campbell (Eds.), *Quasi-experimentation*. Boston, MA: Houghton Mifflin.

Reichardt, C. S., & Mark, M. M. (1998). Quasi-experimentation. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods*. Thousand Oaks, CA: Sage Publications.

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

Reichardt, C. S., Trochim, W. M. K., & Cappelleri, J. C. (1995). Reports of the death of regression-discontinuity analysis are greatly exaggerated. *Evaluation Review, 19*, 39–63.

Renger, R., & Hurley, C. (2006). From theory to practice: Lessons learned in the application of the ATM approach to developing logic models. *Evaluation and Program Planning, 29*, 106–119.

Renger, R., & Titcomb, A. (2002). A three-step approach to teaching logic models. *American Journal of Evaluation, 23*, 493–503.

Rezmovic, E. L., Cook, T. J., & Dobson, L. D. (1981). Beyond random assignment: Factors affecting evaluation integrity. *Evaluation Review, 5*, 51–67.

Ricciardi, T. N., & Tabb, K. (2006). Role of information technology in measuring and improving quality. In D. B. Nash & N. I. Neil (Eds.), *The quality solution: The stakeholder's guide to improving health care*. Boston, MA: Jones and Bartlett Publishers.

Rich, E. C., Gifford, G., Luxenberg, M., & Dowd, B. (1990). The relationship of house staff experience to the cost and quality of inpatient care. *JAMA, 263*, 953–957.

Riecken, H. W., & Boruch, R. F. (Eds.). (1974). *Social experimentation: A method for planning and evaluating social intervention*. New York, NY: Academic Press.

Ripsin, C. M., Keenen, J. M., Jacobs, D. R., Jr., Elmer, P. J., Welch, R. R., Van Horn, L., … Kestin, M. (1992). Oat products and lipid lowering. A meta-analysis. *JAMA, 267*, 3317–3325.

Rivlin, A. M. (1990, October). *Evaluation and public policy*. Invited address presented at the meeting of the American evaluation association. Washington DC.

Robinson, E. A. R., & Doueck, H. J. (1994). Implications of the pre/post/then design for evaluating social group work. *Research on Social Work Practice, 4*, 224–239.

Rog, D. J. (2015). Infusing theory into practice, practice into theory. *American Journal of Evaluation, 36*, 223–238. doi:10.1177/1098214015573068

Rogers-Dillon, R. (2004). *The welfare experiments: Politics and policy evaluation*. Stanford, CA: Stanford University Press.

Rook, K. S. (1987). Effects of case history versus abstract information on health attitudes and behavior. *Journal of Applied Social Psychology, 17*, 533–553.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45*, 775–778.

Ross, J. A., Barkaoui, K., & Scott, G. (2007). Evaluations that consider the cost of educational programs: The contribution of high-quality studies. *American Journal of Evaluation, 28*, 477–492.

Rossi, P. H. (1978). Issues in the evaluation of human services delivery. *Evaluation Quarterly, 2*, 573–599.

Rossman, G. B., & Wilson, B. L. (1985). Numbers and words: Combining quantitative and qualitative methods in a single large-scale evaluation study. *Evaluation Review, 9*, 627–644.

Roth, J. (1990). Needs and the needs assessment process (reprinted from 1978). *Evaluation Practice, 11*, 141–143.

Sackett, P. R., & Mullen, E. J. (1993). Beyond formal experimental design: Towards an expanded view of the training evaluation process. *Personnel Psychology, 46*, 613–627.

Sanchez, J. R., & Laanan, F. S. (1997). The economic returns of a community college education: ERIC review. *Community College Review, 25*(Winter), 73–87.

Sarason, S. B. (1972). *The creation of settings and the future societies*, San Francisco, CA: Jossey-Bass.

Saunders, R. P., Ward, D., Felton, G. M., Dowda, M., & Pate, R. R. (2006). Examining the link between program implementation and behavior outcomes in the lifestyle for activity program (LEAP). *Evaluation and Program Planning, 29*, 352–364.

Savaya, R. (1998). The under-use of psychological services by Israeli Arabs: An examination of the roles of negative attitudes and the use of alternative sources of help. *International Social Work, 41*, 195–209.

Schermerhorn, J. R., Jr., Hunt, J. G., & Hunt, R. N. (2008). *Organizational behavior* (10th ed.). Hoboken, NJ: Wiley.

Schneider, M. J., Chapman, D. D., & Voth, D. E. (1985). Senior center participation: A two-stage approach to impact evaluation. *The Gerontologist, 25*, 194–200.

Schneider, A. L., & Darcy, R. E. (1984). Policy implications of using significance tests in evaluation research. *Evaluation Review, 8*, 573–582.

Schwandt, T. A. (2014). On the mutually informing relationship between practice and theory in evaluation. *American Journal of Evaluation, 35*, 231–236. doi:10.1177/1098214013503703

Schwartz, R., & Mayne, J. (2005). Assuring the quality of evaluative information: Theory and practice. *Evaluation and Program Planning, 28*, 1–14.

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *American Journal of Evaluation, 22*, 127–160.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation*. Chicago, IL: Rand-McNally.

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

Scriven, M. (1997a). Truth and objectivity in evaluation. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook*. Thousand Oaks, CA: Sage Publications.

Scriven, M. (2003). Evaluation in the new millennium: The transdisciplinary vision. In S. I. Donaldson & M. Scriven (Eds.), *Evaluating social programs and problems: Visions for the new millennium*. Mahwah, NJ: Erlbaum.

Scriven, M. (2007). Key evaluation checklist. *The Evaluation Center*. Retrieved from http://www.wmich.edu/sites/default/files/attachments/u350/2014/key%20evaluation%20checklist.pdf.

Scriven, M., & Roth, J. (1990). Special feature: Needs assessment (reprinted from 1976). *Evaluation Practice, 11*, 135–140.

Seidling, M. B. (2015). Evaluator certification and credentialing revisited: A survey of American Evaluation Association members in the United States. In J. W. Altschuld & M. Engle (Eds.), *New Directions for Evaluation: No. 145. Accreditation, certification, and credentialing: Relevant concerns for U.S. evaluators*. San Francisco, CA: Jossey-Bass.

Seligman, C., & Finegan, J. E. (1990). A two-factor model of energy and water conservation. In J. Edwards, R. S. Tindale, L. Heath, & E. J. Posavac (Eds.), *Social influence processes and prevention*. New York, NY: Plenum.

Seligman, M. (2011). *Flourish: A visionary new understanding of happiness and well-being*. New York, NY: Simon and Schuster.

Senn, C. Y., & Forrest, A. (2016). 'And then one night when I went to class…': The impact of sexual assault bystander intervention workshops incorporated in academic courses. *Psychology of Violence, 6*, 607–618.

Shackman, G. (2015). Accreditation, certification, credentialing: Does it help? In J. W. Altschuld & M. Engle (Eds.), *New Directions for Evaluation: No. 145. Accreditation, certification, and credentialing: Relevant concerns for U.S. evaluators*. San Francisco, CA: Jossey-Bass.

Shadish, W. R. (1993). Critical multiplism: A research strategy and its attendant tactics. In L. Sechrest (Ed.), *Program evaluation: A pluralistic enterprise. New directions for program evaluation* (No. 60). San Francisco, CA: Jossey-Bass.

Shadish, W. R. (1995). Philosophy of science and the quantitative-qualitative debates: Thirteen common errors. *Evaluation and Program Planning, 18*, 63–75.

Shadish, W. R. (2002). Revisiting field experimentation: Field notes for the future. *Psychological Methods, 7*, 3–18.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Boston, MA: Houghton Mifflin.

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

Shadish, W. R., Jr., Orwin, R. G., Silber, B. G., & Bootzin, R. R. (1985). The subjective well-being of mental patients in nursing homes. *Evaluation and Program Planning, 8*, 239–250.

Shaughnessy, J. J., Zechmeister, E. B., & Zechmeister, J. S. (2008). *Research methods in psychology* (8th ed.). New York, NY: McGraw-Hill.

Shepard, L. A. (1990). *"Inflating test score gains": Is it old norms or teaching the test?* Los Angeles, CA: UCLA Center for Research on Evaluation, Standards, and Student Teaching, CSE Technical Report 307.

Shipley, R. H. (1976). Effects of companion program on college student volunteers and mental patients. *Journal of Consulting and Clinical Psychology, 4*, 688–689.

Shrauger, J. S., & Osberg, T. M. (1981). The relative accuracy of self-predictions and judgments by others in psychological assessment. *Psychological Bulletin, 90*, 322–351.

Shulha, L. M., Whitmore, E., Cousins, J. B., Gilbert, N., & al Hudib, H. (2016). Introducing evidence-based principles to guide collaborative approaches to evaluation. *American Journal of Evaluation, 37*, 193–215. doi:10.1177/1098214015615230

Sieber, J. E. (1998). Planning ethically responsible research. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods*. Thousand Oaks, CA: Sage Publications.

Sieber, S. D. (1981). *Fatal remedies: The ironies of social intervention*. New York, NY: Plenum.

Silverman, M., Ricci, E. M., & Gunter, M. J. (1990). Strategies for increasing the rigor of qualitative methods in evaluation of health care programs. *Evaluation Review, 14*, 57–74.

Silverman, W. A. (1977, June). The lesson of retrolental fibroplasia. *Scientific American, 236*, 100–107.

Sirotnik, K. A. (Ed.). (1990). *Evaluation and social justice*. San Francisco, CA: Jossey-Bass.

Slovic, P. (1993). Perceived risk, trust, and democracy. *Risk Analysis, 13*, 675–682.

Smith, M. F. (1989). *Evaluability assessment: A practical approach*. Boston, MA: Kluwer Academic Publishers.

Solberg, L. I., Reger, L. A., Pearson, T. L., Cherney, L. M., O'Connor, P. J., Freeman, S. L., … Bishop, D. B. (1997). Using continuous quality improvement to improve diabetes care in populations. *Journal of Quality Improvement, 23*, 581–592.

Solmeyer, A. R., & Constance, N. (2015). Unpacking the "Black Box" of social programs and policies. *American Journal of Evaluation, 36*, 470–474. doi:10.1177/1098214015600786

Sommer, R. (1990). Local research. *The Journal of Social Issues, 46*(Spring), 203–214.

Sonnichsen, R. C. (1994). Evaluators as change agents. In J. S. Wholey, H. Hatry, & K. Newcomer (Eds.), *Handbook of practical program evaluation*. San Francisco, CA: Jossey Bass.

Sonnichsen, R. C. (2000). *High impact internal evaluation: A practitioner's guide to evaluating and consulting inside organizations*. Thousand Oaks, CA: Sage Publications.

Spiring, F. A. (1994). A bill's effect on alcohol-related traffic fatalities. *Quality Progress, 27*(2), 35–38.

Spiro, S. E., Shalev, A., Solomon, Z., & Kotler, M. (1989). Self-reported change versus changed self-report: Contradictory findings of an evaluation of a treatment program for war veterans suffering from post-traumatic stress disorder. *Evaluation Review, 13*, 533–549.

St. Clair, T., Cook, T. D., & Hallberg, K. (2014). Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. *American Journal of Evaluation, 35*, 331–327. doi:10.1177/1098214014527337

Steele, S. (1990). *The content of our character*. New York, NY: St. Martin's Press.

Stein, S., & Recktenwald, W. (1990, November 11). City parks are no place to play. *Chicago Tribune*, Sec. 1, pp. 1, 18.

Stevahn, L., King, J. A., Ghere, G., & Minnema, J. (2005). Establishing essential competencies for program evaluators. *American Journal of Evaluation, 26*, 43–59.

Stricker, G., & Trierweiler, S. J. (1995). The local clinical scientist: A bridge between science and practice. *American Psychologist, 50*, 995–1002. doi:10.1037/0003-066X.50.12.995

Stucker, C. (2004). *Mystery shopper's manual* (6th ed.). Sugar Land, TX: Special Interests Publishing.

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

Stufflebeam, D. L., & Coryn, C. L. S. (2014). *Evaluation theory, models and applications*. San Francisco, CA: Jossey-Bass.

Sturges, K. M., & Howley, C. (2017). Responsive meta-evaluation. *American Journal of Evaluation, 38*, 126–137. doi:10.1177/1098214016630405

Sturges, K. M. (2015). Complicity revisited. *American Journal of Evaluation, 36*, 461–469. doi:10.1177/1098214015583329

Sullivan, J. M., & Snowden, L. R. (1981). Monitoring frequency of client problems. *Evaluation Review, 5*, 822–833.

Summerfelt, W. T. (2003). Program strength and fidelity in evaluation. *Applied Developmental Science, 7*, 55–61.

Sussna, E., & Heinemann, H. N. (1972). The education of health manpower in a two-year college: An evaluation model. *Socio-Economic Planning Science, 6*, 21–30.

Svensson, K. (1997). The analysis of evaluation of foreign aid. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook*. Thousand Oaks, CA: Sage Publications.

Swain, J. F., Rouse, I. L., Curley, C. B., & Sacks, F. M. (1990). Comparison of the effects of oat bran and low-fiber wheat on serum lipoprotein levels and blood pressure. *New England Journal of Medicine, 322*, 147–152.

Tamura, M. K., Covinsky, K. E., Chertow, G. M., Yaffe, K., Landefeld, C. S., & McCulloch, C. E. (2009). Functional status of elderly adults before and after initiation of dialysis. *New England Journal of Medicine, 361*, 1539–1547.

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

Taylor, P. J., Russ-Eft, D. F., & Taylor, H. (2009). Guilding the outcome by tarnishing the past: Inflationary biases in retrospective pretests. *American Journal of Evaluation, 30*, 31–43.

Testa, M. A., Anderson, R. B., Nackley, J. F., & Hollenberg, N. K. (1993). Quality of life and antihypertensive therapy of men. *New England Journal of Medicine, 328*, 907–913.

The trouble with dependent variables. (1990, Spring). *Dialogue: Society for Personality and Social Psychology*, p. 9.

Thomas, V. G., & Parsons, B. A. (2017). Culturally responsive evaluation meets systems-oriented evaluation. *American Journal of Evaluation, 38*, 7–28. doi:10.1177/1098214016644069

Torres, R. T. (1994). Linking individual and organizational learning: The internalization and externalization of evaluation. *Evaluation and Program Planning, 17*, 327–338.

Torres, R. T., Preskill, H. S., & Piontek, M. E. (2004). *Evaluation strategies for communicating and reporting: Enhancing learning in organizations* (Rev. ed.). Thousand Oaks, CA: Sage Publications.

Trochim, W. M. K. (1984). *Research design for program evaluation: The regression discontinuity approach*. Beverly Hills, CA: Sage Publications.

Trochim, W. M. K. (Ed.). (1986). *Advances in quasi-experimental design and analysis*. San Francisco, CA: Jossey-Bass.

Trochim, W. M. K. (1990). The regression-discontinuity design. In Research methodology: Strengthening causal interpretations of nonexperimental data, ed. L. Sechrest, E. Perrin, & J. Bunker. Rockville, MD: U.S. Department of Health and Human Services, Agency for Health Care Policy and Research, (PHS) 90 3454.

Turner, G. (1999). Peer support and young people's health. *Journal of Adolescence, 22*, 567–572.

Turner, M. A., & Zimmerman, W. (1994). Acting for the sake of research: The use of role-playing in evaluation. In J. S. Wholey, H. R. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation*. San Francisco, CA: Jossey-Bass.

Tyson, T. J. (1985). The evaluation and monitoring of a Medicaid second surgical opinion program.

Evaluation and Program Planning, 8, 207–216.

University of California at Davis (2014). *Afghanistan Agricultural Extension Project II (AAEP II), Final Report*, Retrieved from https://pdf.usaid.gov/pdf_docs/PA00SSVM.pdf.

U.S. Department of Health and Human Services. (n.d.). Code of Federal Regulations, 45 CFR 46. *Protection of Human Subjects*. Retrieved from https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html.

Van den Eynde, J., Veno, A., & Hart, A. (2003). They look good but don't work: A case study of global performance indicators in crime prevention. *Evaluation and Program Planning, 26*, 237–248.

Vermillion, J. M., & Pfeiffer, S. I. (1993). Treatment outcome and continuous quality improvement: Two aspects of program evaluation. *The Psychiatric Hospital, 24*, 9–14.

Viscusi, W. K. (1992). *Fatal tradeoffs: Public and private responsibilities for risk*. New York, NY: Oxford University Press.

Vojtecky, M. A., & Schmitz, M. F. (1986). Program evaluation and health and safety training. *Journal of Safety Research, 17*, 57–63.

Vroom, P. I., Colombo, M., & Nahan, N. (1994). Confronting ideology and self-interest: Misuse of evaluation. In C. J. Stevens & M. Dial (Ed.), *Preventing the misuse of evaluation. New directions for program evaluation* (No. 64). San Francisco, CA: Jossey-Bass.

W. K. Kellogg Foundation. (2004). *Logic model development guide: Using logic models to bring together planning, evaluation, and action*. Retrieved from https://www.wkkf.org/resource-directory/resource/2006/02/wk-kellogg-foundation-logic-model-development-guide.

W. K. Kellogg Foundation. (2010). *Evaluation handbook*. Retrieved from https://www.wkkf.org/resource-directory/resource/2010/w-k-kellogg-foundation-evaluation-handbook

Waller, B. (2005). *Consider ethics: Theory, readings and contemporary issues*. New York, NY: Longman.

Walvoord, B. E. (2004). *Assessment clear and simple*. San Francisco, CA: Jossey-Bass.

Wandersman, A., Alia, K., Cook, B. S., Hsu, L. L., & Ramaswamy, R. (2016). Evidence-based interventions are necessary but not sufficient for achieving outcomes in each setting in a complex world. *American Journal of Evaluation, 37*, 544–561. doi:10.1177/1098214016660613

Washburn, G. (2001, July 5). 2nd probe also finds cabs shun disabled. *Chicago Tribune*, Sec. 2, pp. 1, 2.

Watkins, R., & Kavale, J. (2014). Needs: Defining what you are assessing. In J. W. Altschuld & R. Watkins (Eds.), *New Directions for Evaluation: No. 144. Needs assessment: Trends and a view toward the future*. San Francisco, CA: Jossey-Bass.

Webb, E. J., Campbell, D. T., Schwartz, R. D., Sechrest, L., & Grove, J. B. (1981). *Nonreactive measures in the social sciences* (2nd ed.). Boston, MA: Houghton-Mifflin.

Wedman, J. (2014). Needs assessments in the private sector. In J. W. Altschuld & R.Watkins (Eds.), *New Directions for Evaluation: No. 144. Needs assessment: Trends and a view toward the future*. San Francisco, CA: Jossey-Bass.

Weinreich, N. K. (1999). *Hands-on social marketing: A step-by-step guide*. Thousand Oaks, CA: Sage Publications.

Weiss, C. H. (1988). If program decisions hinged only on information: A response to Patton. *Evaluation Practice, 9*(3), 15–28.

Weiss, C. H. (1998). Have we learned anything new about the use of evaluation? *American Journal of Evaluation, 19*, 21–33.

Weiss, C. H. (2002). What to do until the random assigner comes. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education*. Washington, DC: Brookings Institution Press.

Weiss, C. H., Murphy-Graham, E., Petrosini, A., & Gandhi, A. G. (2008). The fairy godmother—and her warts: Making the dream of evidence-based policy come true. *American Journal of Evaluation, 29*, 29–47.

Weitzman, E. A., & Miles, M. B. (1995). *Computer programs for qualitative data analysis: A software sourcebook*. Thousand Oaks, CA: Sage Publications.

Wholey, J. S. (1991, September). Using program evaluation to improve program performance. *The Bureaucrat*, pp. 55–59.

Wholey, J. S. (1999). Quality control: Assessing the accuracy and usefulness of performance measurement systems. In H. P. Hatry (Ed.), *Performance measurement: Getting results* (pp. 217–239). Washington, DC: The Urban Institute.

Wholey, J. S. (2004). Evaluability assessment. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (Rev. ed.). San Francisco, CA: Jossey Bass.

Williams, S. C., Schmaltz, S. P., Morton, D. J., Koss, R. G., & Loeb, J. M. (2005). Quality of care in U.S. hospitals as reflected by standardized measures, 2002–2004. *The New England Journal of Medicine, 353*(3), 255–264.

Winett, R. A. (1995). A framework for health promotion and disease prevention programs. *American Psychologist, 50*, 341–350.

Witkin, B. R., & Altschuld, J. W. (1995). *Planning and conducting needs assessments: A practical guide* (Rev. ed.). Thousand Oaks, CA: Sage Publications.

Yampolskaya, S., Nesman, T. M., Hernandez, M., & Koch, D. (2004). Using concept mapping to develop a logic model and articulate a program theory: A case example. *American Journal of Evaluation, 25*, 191–207.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Yates, B. T. (1994). Toward the incorporation of costs, cost-effectiveness analysis, and cost-benefit analysis into clinical research. *Journal of Consulting and Clinical Psychology, 62*, 729–736.

Yates, B. T. (1999). *Measuring and improving costs, cost-effectiveness, and cost-benefit for substance abuse treatment programs* (NIH Publication Number 99-4518). Bethesda, MD: National Institute on Drug Abuse.

Yeaton, W. H., & Sechrest, L. (1986). Use and misuse of no-difference findings in eliminating threats to validity. *Evaluation Review, 10*, 836–852.

Yeaton, W. H., & Sechrest, L. (1987). Assessing factors influencing acceptance of no-difference research. *Evaluation Review, 11*, 131–142.

Yeh, S. S. (2006). *Raising student achievement through rapid assessment and test reform*. New York, NY: Teachers College Press.

Yeh, S. S. (2009). Shifting the bell curve: The benefits and costs of raising student achievement. *Evaluation and Program Planning, 32*, 74–82.

Zechmeister, E. B., & Posavac, E. J. (2003). *Data analysis and interpretation in the behavioral sciences*. Belmont, CA: Wadsworth.

Zigler, E., & Trickett, P. K. (1978). IQ, social competence, and evaluation of early childhood intervention programs. *American Psychologist, 33*, 789–798.

Author Index

Abelson, R. P., 232, 269, 293
Abry, T., 8, 293
Abt, C. C., 296
Ackoff, R. L., 148, 293
Adams, M. E., 78, 294
Adams, P., 145, 308
Aiken, L. S., 226, 293
Akins, S. A., 104, 306
al Hudib, H., 311
Alia, K., 60, 314
Alkin, M. C., 16, 43, 313
Allen, C., 223, 293
Altman, D. G., 136, 296
Altschuld, J. W., 56, 125, 126, 138, 142, 293, 302, 303, 304, 310, 311, 314, 315
Anderson, C. A., 200, 293
Anderson, E. A., 95
Anderson, R. B., 136, 313
Anderson, S., 50, 307
Angelo, T. A., 108, 293
Appleton, H., 296
Argyris, C., 282, 293
Aronson, E., 296
Arvey, R. D., 190, 294
Aslanian, C. B., 136, 296
Aspelmeier, J., 206, 294
Atkinson, C. M., 115, 294
Atman, C. J., 123, 306
Auld, G., 85, 300
Averch, H. A., 81, 294
Avery, D. J., 246, 301
Ayers, J., 89, 298
Azzam, T., 15, 299

Babbie, E., 97, 129, 131, 294
Baldwin, K., 212, 305
Ball, S., 115, 294
Bamberger, E., 230, 302

Banks, S. M., 136, 295
Banta, T. W., 307
Barkaoui, K., 238, 309
Bartlett, S., 209, 302
Bauer, S. M., 33, 294
Beaumont, T., 84, 300
Belemvire, A., 308
Bennett, C. A., 299
Benson, P. G., 187, 304
Bentham, J., 115
Berk, R. A., 85, 294
Bickman, L., 35, 101, 294, 299, 303, 305, 309, 311
Bingham, S. G., 114, 294
Birkby, B., 22, 23
Bishop, D. B., 312
Bogartz, G. A., 115, 294
Boggs, A., 134, 303
Boisen, L., 28, 302
Boles, S. M., 299
Boone, S., 297
Bootzin, R. R., 78, 311
Boruch, R. F., 10, 213, 224, 276, 294, 303, 309, 314
Bostrom, A., 123, 306
Botcheva, L., 43, 294
Bounthavong, M., 257, 306
Bours, D., 295, 298
Bowen, A. R., 45, 302
Bowen, M. L., 89, 109, 117, 294
Boxmeyer, C. L., 279, 299
Braizerman, M., 296
Branchey, L., 299
Brandon, P. R., 5, 294
Branson, M. P., 95
Brewer, M. B., 226, 296
Brightwall, D. R., 299
Brinkerhoff, R. O., 79, 168, 294
Brockway, J. H., 262, 300
Bronson, P., 105, 294
Brotherton, S. E., 35, 50, 308
Brotman, B., 108, 294
Brown, A. L., 28, 302
Brown, J. B., 78, 294
Brown, R. D., 115, 306
Browne, G., 224, 305
Bryant, F. B., 191, 262, 294, 298, 300, 305, 307
Bryk, A. S., 32, 294
Bunda, M. A., 115, 295

Bunker, J., 297
Bushway, S., 243, 269, 295
Bybee, D., 144, 306

Cagle, L. T., 136, 295
Caidin, M., 187, 295
Cambron, S., 100
Campbell, D. T., 18, 86, 89, 105, 148, 149, 167, 177, 185, 198, 202, 213, 219, 295, 296, 309, 311, 314
Cannell, J. J., 58, 295
Card, J. J., 48, 295
Carey, R. G., 26, 110, 200, 224, 295, 308
Carley, M., 127, 295
Carmichael, D. H., 95
Carroll, J. L., 226, 293
Caruthers, F. A., 15, 315
Casey, M. A., 93, 132, 302
Castelin, K., 273
Castner, L., 136, 295
Caulkins, J. P., 243, 269, 295
Cervenka, C., 96, 305
Chaintreau, A., 84, 299
Chalkley, M. A., 60, 303
Chalmers, T. C., 269, 299
Champney, T. F., 285, 295
Chan, K-C., 135, 295
Chapman, D. D., 242, 310
Chelimsky, E., 14, 295, 310, 313
Chen, M., 162, 306
Cherney, L. M., 312
Chertow, G. M., 313
Chiesa, J., 243, 269, 295
Chong, J., 89, 295
Chou, T., 305
Christie, C. A., 283, 299
Cicarelli, V. G., 60, 296
Clark, C., 250, 296
Cobb, P., 50, 306
Cody, S., 136, 295
Coghlan, A. T., 18, 168, 278, 308
Cohen, J., 183, 229, 296
Cole, D. A., 190, 294
Colombo, M., 147, 314
Comer, J. S., 305
Conner, R. F., 136, 296
Constance, N., 8, 27, 312
Cook, B. S., 60, 314
Cook, T. C., 225, 296
Cook, T. D., 14, 35, 41, 42, 50, 53, 62, 68, 80, 105, 115, 149, 156, 157, 185, 198, 202, 203, 213, 219, 296, 309, 311, 312

Cook, T. J., 227, 309
Cooper, W. H., 60, 296
Corbin, J. M., 96, 156, 296
Cordray, D. S., 50, 214, 278, 296, 303, 307
Cornacchio, D., 305
Coryn, C. L. S., 26, 312
Cosenza, C., 101, 299
Cousins, J. B., 5, 299, 311
Covinsky, K. E., 313
Cracknell, B., 278, 299
Crano, W. D., 226, 296
Creer, T. L., 167, 305
Crompton, D. W., 280, 296
Cross, K. P., 108, 293
Crosse, S., 35, 110, 111, 113, 144, 225, 277, 303
Crum, K. A., 298
Cumming, G., 266, 296
Cummings, S. R., 253, 296
Curley, C. B., 113, 313

Dao, A., 298
Dao, T. N., 298
Darcy, R. E., 113
Datta, L., 107, 169, 170, 227, 296, 297
Davidshofer, C. O., 86, 88, 94, 188, 306
Dawes, R. M., 6, 43, 67, 223, 297
de Groot, D., 308
de Melker, H. E., 5, 300
Dean, P. J., 95
Deck, S., 288, 290
Decker, P. J., 4, 297
Defriese, G. B., 50, 297
Del Mar, C. B., 223, 293
Deming, W. E., 17, 148, 223, 297
Dennis, M. L., 224, 227, 228, 297, 300
Denzin, N. K., 93, 170, 297
Derby, A., 252, 297
Derman, R. M., 299
DeSerisy, M., 305
Devarajan, S., 135, 297
Dever, M., 153
Devine, E. C., 53, 296
Dial, M., 226, 297, 314
Dickie, I., 233, 235, 297
Dietz, T., 308
Dobson, L. D., 227, 309
Dollar, D., 135, 297
Donaldson, S. I., 15, 18, 33, 297, 302, 310

Doueck, H. J., 191, 309
Dowd, B., 204, 309
Dowda, M., 112, 310
Drexler, T., 196, 298
Duncan, B., 217, 297
Duncan, S. H., 235, 297
Dunkle, J., 35, 110, 111, 113, 144, 225, 277, 303
Dunning, D., 41, 298
Durand, R., 4, 297
Dyrbye, L. N., 71, 297
Dyson, L., 7, 52, 299

Easterly, W., 135, 297
Eckert, P. S., 212, 305
Eddy, D. M., 194, 223, 297
Edmondson, A. C., 148, 283, 297
Edwards, J. D., 298, 305, 307, 310
Egan, G., 138, 298
Eisenberg, L., 182, 298
Eliot, S., 162, 306
Elmer, P. J., 309
Emrick, C. D., 299
Engle, M., 302, 304, 310, 311
Epley, N., 41, 298
Epstein, I., 276, 307
Etzioni, A., 298
Evans, C., 205, 304
Evans, R. G., 249, 298
Evans, R. I., 52, 298
Everingham, S. S., 243, 269, 295
Excell, S., 298

Fals-Stewart, W., 238, 298
Faulkner, L., 89, 298
Felson, M., 127, 298
Felson, R. B., 6, 298
Felton, G. M., 112, 310
Fendrick, A. M., 252, 301
Fernandez, J. S., 44, 45, 302
Ferrari, J. R., 196, 298
Ferriss, A. L., 127, 298
Fetterman, D. M., 9, 45, 158, 298, 302
Figueredo, A. J., 284, 298
Finch, S., 266, 296
Fine, M., 126, 298
Finegan, J. E., 52, 310
Fischhoff, B., 123, 306
Fitch, R. W., 281, 298

Fitzpatrick, J., 70, 298
Fleischer, D. N., 283, 299
Forrest, A., 28, 311
Fountain, D., 299
Fowler, F. J., Jr., 101, 299
Franklin, M., 102
Freeman, S. L., 312
Freiman, J. A., 269, 299
French, M. T., 243, 244, 248, 299, 308
Fross, K., 278, 299
Fukunaga, L. L., 5, 294
Fuller, R. K., 228, 299

Gabielkov, M., 84, 299
Gagne, R. M., 310
Galport, N., 15, 299
Gandhi, A. G., 65, 194, 276, 314
Garland, A. F., 279, 299
Gates, E., 7, 52, 299
Gehrs, M., 145, 299
Gero, A., 89, 295
Gerstein, D. R., 248, 299
Ghere, G., 15, 302, 312
Gifford, G., 204, 309
Gilbert, J. P., 224, 299
Gilbert, N., 5, 311
Gill, S. J., 282, 292, 299
Gillmore, G. M., 149, 299
Ginsburg, G. S., 95
Glasgow, R. E., 52, 299
Glasziou, P., 223, 293
Glenmullen, J., 66, 299
Goering, P. N., 285, 299
Goggin, W. C., 95
Goldman, M., 84, 300
Golik, A., 305
Gooler, L. E., 18, 33, 297
Gorman, P., 162, 306
Gottfredson, D., 145, 303
Granger, R. L., 60, 296
Grasso, A. J., 276, 307
Green, J., 81, 300
Green, R. S., 82, 300
Greene, J. C., 170, 300
Greeno, C., 48, 295
Greenwald, A. G., 149, 299
Greiner, J. M., 82, 93, 300
Gross, D. M., 242, 300

Grove, J. B., 86, 314
Gruenewald, P. J., 230, 302
Guba, E. G., 82, 164, 167, 169, 170, 300, 303
Guest, G., 162, 306
Guilbault, R. L., 191, 262, 294, 300
Gunter, M. J., 167, 311

Hacsi, T. A., 294, 296
Hahn, C. K., 28, 300
Haine, R. A., 279, 299
Hallberg, K., 203, 312
Harmsen, I. A., 5, 300
Harper, D., 224, 305
Hart, A., 90, 313
Hartog, J., 243, 300
Hartog, S. B., 190, 306
Hartung, B. M., 92, 145, 308
Harwood, H. J., 299
Hatry, H. P., 19, 80, 300, 301, 312, 315
Hatry, H. R., 294, 297, 300, 313
Health, L., 298
Heath, C., 123, 131, 300
Heath, D., 131, 300
Heath, L., 300, 305, 307, 310
Hedrick, T. E., 170, 300
Hegarty, T. W., 258, 300
Heinemann, H. N., 243, 312
Helford, M., 35, 50, 308
Hemphill, J. F., 95
Henderson, E., 298, 305, 307
Henderson, G. R., 200, 300
Henderson, M., 85, 300
Hendricks, M., 261, 262, 301
Herman, P. M., 246, 301
Hernandez, M., 50, 315
Hilkevitch, J., 148, 301
Ho, E. A., 132, 308
Hogan, R. R., 79, 301
Hollenberg, N. K., 136, 313
Hollis, J. F., 299
Holmgren, T., 135, 297
Holter, M., 144, 306
Hopson, R. K., 15, 315
Horton, S. V., 203, 301
Hougland, J. G., Jr., 85, 301
House, P. W., 250, 301
Howard, P. K., 114, 226, 301
Howell, A. J., 95

Howley, C., 26, 312
Hsiung, S., 226, 293
Hsu, L. L., 60, 314
Hudib, H., 5
Huebner, T. A., 294, 296
Huffman, L. C., 43, 294
Hulleman, C. S., 8, 293
Hunt, J. G., 13, 310
Hunt, R. N., 13, 310
Huq, S., 89, 298
Hurley, C., 52, 61, 309
Huse, I., 15, 304

Iber, F. L., 299
Impara, J. C., 95
Ioannidis, J. P. A., 112, 301

Jacobi, M., 136, 296
Jacobs, D. R., Jr., 309
Jacobs, G. A., 28, 300
Jacobsen, N. S., 178, 301
Jason, L. A., 82, 301
Jerrell, J. M., 82, 300
Joglekar, P. N., 253, 301
Johnson, N. F., 85, 300
Johnson, R. A., 299
Judd, C. M., 301
Julnes, G., 281, 301

Kaftarian, S. J., 45, 302
Kahn, J., 243, 301
Kane, R. A., 247
Kane, R. L., 247, 301
Karwath, R., 59, 301
Katz, M. M., 82, 301
Kavale, J., 125, 314
Keenen, J. M., 309
Kenny, D. A., 301
Kent, D. M., 252, 301
Kernic, M. A., 245, 302
Kestin, M., 309
Kibel, B. M., 79, 302
King, J., 11, 69, 238, 302
King, J. A., 15, 238, 286, 302, 312
Kingsbury, K., 252, 302
Kirkman, D. M., 4, 297
Kirkpatrick, J., 308
Klerman, J. A., 209, 302
Klitzner, M., 230, 302

Klostermann, K., 238, 298
Kniffley, S., 173
Koch, D., 50, 315
Koelsch, L. D., 28, 302
Kohfeldt, D., 45, 302
Kok, G., 5, 300
Kornbluh, M., 170, 306
Koss, R. G., 145, 315
Kotler, M., 190, 312
Kramer, P. D., 66, 302
Krueger, R. A., 93, 132, 133, 302
Kuebler, R. R., 269, 299
Kuji-Shikatani, K., 15, 302
Kuo, E., 245, 302
Kurtines, W. M., 95

Laanan, F. S., 243, 310
Landefeld, C. S., 313
Langa, K. M., 252, 301
Langhout, R. D., 44, 45, 46, 302
Lave, L., 123, 306
LaVelle, J. M., 15, 302
Lavrakas, P. J., 93, 303
Lazar, I., 60, 113, 303
Le Menestrel, S. M., 17, 303
Lebedun, M., 205, 304
Leeuw, F. L., 35, 303
Legout, A., 84, 299
Leik, R. K., 60, 303
Lepicki, T., 134, 303
Leslie, L. A., 95
Levin, H. M., 11, 13, 246, 248, 303
Leviton, L. C., 14, 56, 60, 80, 149, 156, 276, 296, 303, 311
Lewandowski, L., 144, 306
Liew, S. M., 308
Light, R. J., 167, 224, 226, 299, 303
Lincoln, K. D., 6, 303
Lincoln, Y. S., 18, 82, 93, 164, 167, 169, 170, 297, 300, 303
Lindzey, G., 296
Linfield, K., 118
Linn, R. L., 58, 303
Liotta, R. F., 82, 301
Lipsey, M. W., 35, 36, 50, 87, 90, 110, 112, 113, 144, 145, 181, 225, 229, 266, 268, 269, 277, 278, 281, 303, 304
Lockhart, D. C., 212, 305
Loeb, J. M., 145, 315
Lopreto, T., 308
Louie, J., 209, 304
Love, A. J., 79, 130, 304

Lowe, R. A., 162, 306
Lubin, B., 205, 304
Lumpkin, P. W., 95
Lumsdaine, A. R., 299
Lurigio, A. J., 81, 147, 304
Luxenberg, M., 204, 309

MacRae, D., Jr., 127, 304
Madden, J. S., 56, 304
Mager, R. F., 88, 304
Maiuro, R. D., 302
Malcolm, M. T., 56, 304
Malhotra, N. K., 132, 304
Malloy, K., 299
Mangione, T. W., 92, 304
Manoff, R. K., 135, 136, 304
Margasak, L., 81, 304
Mark, J., 209, 304
Mark, M. M., 170, 203, 304, 309
Markandya, A., 251, 307
Marshall, E., 102, 111
Mashaal, M., 57, 304
Massarad, M. G., 282, 304
Massey, O. T., 85, 304
Mathison, S., 62, 138, 304
Maxwell, G. S., 167, 304
Mayne, J., 14, 310
McCauley, E., 245, 302
McClary, S., 205, 304
McClave, T., 187, 304
McClelland, G. H., 301
McCulloch, C. E., 313
McDavid, J. C., 15, 304
McDowall, M. A., 300
McEwan, P. J., 11, 13, 246, 248, 303
McGarrell, E. F., 62, 107, 305
McGinn, C., 295, 298
McKenna, K., 162, 306
McKillip, J., 96, 124, 201, 212, 305
McKnight, J., 5, 126, 135, 181, 305
McPhail, A., 224, 305
McSweeny, A. J., 167, 305
McWhorter, J. H., 68, 305
Mead, G. E., 300
Mead, L. M., 27, 305
Meehl, P. E., 281, 305
Meier, S. T., 90, 178, 182, 305
Mendez, D., 41, 305

Merryman, A., 105, 294
Mertens, D. M., 76, 305
Mielke, M. B., 17, 304
Miguel, E. M., 84, 305
Miles, M. B., 96, 164, 314
Millar, D., 308
Millenson, M. L., 249, 305
Millsap, R. E., 190, 306
Minnema, J., 15, 302, 312
Mitra, A., 93, 134, 200, 306
Mohr, L. B., 281, 301
Moirs, K., 96, 305
Mollema, L., 5, 300
Moore, T. J., 223, 306
Morgan, D. L., 162, 306
Morgan, M. G., 123, 306
Morris, J. M., 28, 109, 117, 122, 300, 306
Mortenson, E., 298
Morton, D. J., 145, 315
Mosher, C. J., 104, 306
Moskowitz, J. M., 278, 306
Mosteller, F., 66, 224, 299, 306, 314
Mowbray, C. T., 144, 306
Muennig, P., 257, 306
Mullen, E. J., 113, 310
Munter, C., 50, 306
Murphy, K. R., 86, 88, 94, 188, 306
Murphy-Graham, E., 65, 194, 276, 314

Nackley, J. F., 136, 313
Nagel, S. S., 250, 306
Nahan, N., 147, 314
Namey, E., 162, 306
Nash, D. B., 309
Neal, J. W., 170, 306
Neal, Z. P., 170, 306
Neil, N. I., 309
Nelken, Y., 308
Nesman, T. M., 50, 315
Neuhauser, D., 224, 306
Newcomer, K. E., 19, 294, 297, 300, 301, 312, 313, 315
Newman, D. L., 115, 306
Nickerson, R. S., 231, 268, 306
Norcross, J. C., 51, 119, 308

O'Connor, P. J., 312
O'Dell, K. J., 95
O'Donnell, V. L., 312

Obeng, F., 308
Olsho, L. E. W., 209, 302
Orwin, R. G., 78, 311
Osberg, T. M., 78, 311
Oster, G., 253, 296
Overstreet, E. J., 276, 307
Oyserman, D., 78, 129, 310

Palomba, C. A., 307
Pardo, M. A., 298
Parsons, B. A., 26, 313
Pate, R. R., 112, 310
Patton, M. Q., 26, 60, 96, 111, 162, 175, 276, 307
Paul, G. L., 93, 307
Paulos, J. A., 191, 307
Paulussen, T. W., 5, 300
Pawson, R., 307
Pearce, D., 251, 307
Pearson, T. L., 312
Peck, L. R., 27, 305, 307
Pennington, M., 256
Perkins, O., 137, 307
Perrin, E., 297
Peterson, J. L., 48, 295
Peterson, R. D., 246, 307
Petraitis, J., 123, 300
Petrie, C., 145, 303
Petrosini, A., 65, 194, 276, 314
Petrosino, A., 294, 296
Pfeiffer, S. I., 18, 313
Phillips, J., 212, 305
Pierce, T., 206, 294
Pikula, H., 224, 305
Pillemer, D. B., 167, 226, 268, 303
Pion, G. M., 50, 307
Piontek, M. E., 271, 275, 278, 284, 313
Plake, B. S., 95
Pollard, J. A., 35, 51, 110, 111, 113, 144, 225, 277, 303, 304
Posavac, E. J., 26, 35, 50, 66, 78, 92, 108, 130, 145, 180, 183, 232, 261, 262, 268, 281, 298, 300, 305, 307, 308, 310, 315
Posavac, S. S., 132, 308
Pouw, N., 26, 308
Preskill, H. S., 18, 168, 261, 271, 275, 278, 284, 308, 313
Price, A., 84, 308
Price, J., 308
Pringle, P., 295, 298
Prochaska, J. O., 51, 119, 308
Pulford, J., 145, 308

Raines, B. E., 52, 298
Raines, J. C., 194, 308
Rajkumar, A. S., 244, 248, 308
Rallis, S. F., 156, 167, 170, 297, 300, 309
Ramachandran, A., 84, 299
Ramaswamy, R., 60, 314
Ramirez, G. Y., 95
Ramsay, S., 300
Range, L. M., 95
Rawls, J., 308
Ray, M., 17, 308
Raykov, T., 308
Recktenwald, W., 146, 312
Reger, L. A., 312
Reichardt, C. S., 156, 167, 170, 206, 297, 300, 309
Renger, R., 36, 52, 61, 309
Rezmovic, E. L., 227, 309
Rhoads, C., 209, 304
Ricci, E. M., 167, 311
Ricciardi, T. N., 145, 309
Rich, E. C., 204, 309
Riecken, H. W., 213, 309
Rimm-Kaufman, S., 8, 293
Ripsin, C. M., 113, 309
Rivlin, A. M., 137, 309
Roberts, J., 224, 305
Robinson, E. A. R., 191, 309
Robinson, G. C., 249, 298
Rog, D. J., 50, 101, 225, 299, 303, 305, 309, 311
Rogers, P. J., 294, 296
Rogers-Dillon, R., 226, 309
Rook, K. S., 134, 309
Rosenthal, R., 41, 269, 309
Ross, J. A., 238, 309
Rossi, P. H., 56, 85, 294, 298, 309
Rossiter, C., 230, 302
Rossman, G. B., 167, 309
Roth, J., 124, 309, 310
Rourke, S. B., 299
Rouse, I. L., 113, 313
Rubin, S. M., 253, 296
Ruiter, R. C., 5, 300
Russ-Eft, D. F., 191, 313
Rydell, C. P., 243, 269, 295

Sabath, M. J., 62, 107, 305
Sackett, P. R., 113, 310
Sacks, F. M., 113, 313

Samset, K., 278, 299
Sanbonmatsu, D. M., 132, 308
Sanchez, A. L., 305
Sanchez, J. R., 243, 310
Sarason, S. B., 31, 310
Satele, D., 297
Saunders, R. P., 112, 310
Savaya, R., 150, 310
Schemp, C. S., 246, 301
Scherer, L. L., 114, 294
Schermerhorn, J. R., Jr., 13, 310
Schmaltz, S. P., 145, 315
Schmitz, M. F., 111, 314
Schneider, A. L., 113
Schneider, M. J., 242, 310
Schwalm, D. E., 226, 293
Schwandt, T. A., 56, 310
Schwartz, R. D., 14, 86, 310, 314
Schwarz, N., 78, 129, 310
Scott, A. G., 303
Scott, G., 238, 309
Scott, S., 242, 300
Scriven, M., 3, 13, 17, 18, 23, 33, 65, 68, 124, 287, 297, 310
Sechrest, L., 86, 281, 297, 298, 303, 311, 314, 315
Seidling, M. B., 15, 310
Seligman, C., 52, 310
Seligman, M., 71, 311
Senn, C. Y., 28, 311
Shackman, G., 15, 311
Shadish, W. R., 14, 41, 42, 78, 80, 85, 112, 149, 156, 170, 198, 213, 219, 269, 281, 284, 295, 296, 310, 311, 313
Shaffer, A., 296
Shalev, A., 190, 312
Shanafelt, T. D., 297
Shaughnessy, J. J., 90, 311
Shaw, S., 299
Shekell, C., 50, 306
Shepard, L. A., 58, 311
Sheppard, T. L., 138, 140
Sheridan, J., 145, 308
Shipley, R. H., 85, 311
Shotland, R. L., 170, 304
Shrauger, J. S., 78, 311
Shulha, L. M., 5, 15, 311, 315
Shull, R. D., 250, 301
Sieber, J. E., 5, 32, 311
Sieber, S. D., 54, 114, 226, 311
Silber, B. G., 78, 311

Silverman, M., 167, 311
Silverman, W. A., 223, 311
Silverman, W. K., 95
Sinacore, J. M., 35, 50, 268, 308
Sincich, T., 187, 304
Sirotnik, K. A., 68, 311
Skarr, J., 196, 298
Sloan, J., 297
Slovic, P., 123, 311
Smith, H., Jr., 269, 299
Smith, M., 299
Smith, M. F., 177, 311
Smith-Fowler, H., 299
Snowden, L. R., 85, 312
Solberg, L. I., 145, 312
Solmeyer, A. R., 8, 27, 312
Solomon, Z., 190, 312
Sommer, R., 277, 312
Sonnichsen, R. C., 70, 277, 285, 312
Spiring, F. A., 200, 312
Spiro, S. E., 190, 312
Sporn, D. L., 258, 300
St. Clair, T., 203, 312
Stanley, J. C., 177, 295
Steele, S., 68, 312
Stein, S., 146
Stevahn, L., 15, 286, 302, 312
Stevens, C. J., 297, 314
Stevens, G., 312
Stobart, G., 35, 110, 111, 113, 144, 225, 277, 303
Strauss, A. L., 96, 156, 296
Stricker, G., 312
Strycker, L. A., 299
Stucker, C., 160, 312
Stufflebeam, D. L., 26, 66, 312
Sturges, K. M., 26, 109, 117, 312
Suarez-Balcazar, Y., 298, 305, 307
Sullivan, J. M., 85, 312
Summerfelt, W. T., 112, 312
Sussna, E., 243, 312
Suter, N., 299
Svensson, K., 14, 313
Swain, J. F., 113, 313
Swartz, J., 81, 147, 304

Tabb, K., 145, 309
Tamkin, G., 296
Tamura, M. K., 247, 313

Tan, L., 297
Taut, S., 16, 43, 313
Taylor, H., 191, 313
Taylor, P. J., 191, 313
Terborg, J. R., 299
Testa, M. A., 136, 313
Thomas, V. G., 26, 313
Tilley, N., 307
Tillman, A. S., 89, 109, 117, 294
Tindale, R. S., 298, 305, 307, 310
Titcomb, A., 36, 309
Toms, K., 33, 294
Torres, R. T., 61, 271, 275, 278, 284, 313
Treichel, P., 89, 295
Trickett, P. K., 85, 315
Trierweiler, S. J., 312
Trochim, W. M. K., 209, 211, 313
Truax, P., 178, 301
Turner, G., 52, 95, 313
Turner, M. A., 82, 313
Turner, N. H., 95
Turpin, R. S., 35, 50, 308
Tyler, R. W., 310
Tyson, T. J., 212, 313

Van den Eynde, J., 90, 313
van Dijke, M. L., 300
Van Horn, L., 309
Vander Stoep, A., 245, 302
Vanderslice, V., 126, 298
VanDyke, E., 170, 306
Veno, A., 90, 313
Vermillion, J. M., 18, 313
Viscusi, W. K., 130, 251, 252, 313
Vojtecky, M. A., 111, 314
Voth, D. E., 242, 310
Vroom, P. I., 147, 314

Walahoski, J. S., 17, 303
Waller, B., 115, 314
Walsh, M. E., 246, 301
Walvoord, B. E., 88, 314
Wandersman, A., 9, 45, 60, 158, 298, 302, 314
Ward, D., 112, 310
Warner, K. E., 41, 305
Warren, W. L., 82, 301
Washburn, G., 82, 314
Wasylenki, D. A., 285, 299

Watkins, R., 125, 126, 138, 293, 303, 314
Watt, B., 205, 304
Weaver, G. D., 95
Webb, E. J., 86, 314
Webber, S. J., 296
Wedman, J., 134, 314
Weems, C. F., 95
Weinreich, N. K., 137, 314
Weisburd, D., 145, 303
Weiss, C. H., 65, 123, 144, 194, 269, 276, 280, 314
Weitzman, E. A., 96, 164, 314
Welch, R. R., 309
West, C. P., 297
West, S. G., 226, 293
White, C. R., 43, 294
Whitmore, E., 5, 311
Wholey, J. S., 11, 13, 14, 19, 35, 294, 297, 300, 301, 312, 313, 314, 315
Williams, A. E., 56, 304
Williams, L., 312
Williams, S. C., 145, 315
Wilson, A. T., 76, 305
Wilson, B. L., 167, 309
Wilson, D. B., 181, 229, 268, 304
Winett, R. A., 136, 315
Wintfeld, N., 81, 300
Witkin, B. R., 56, 293, 315
Wohlfarth, D., 118, 120
Wu, L., 85, 304

Yaffe, K., 313
Yampolskaya, S., 50, 315
Yarbrough, D. B., 15, 315
Yates, B. T., 238, 249, 298, 315
Yeaton, W. H., 281, 315
Yeh, S. S., 243, 244, 315

Zaal, F., 308
Zechmeister, E. B., 90, 180, 183, 311, 315
Zechmeister, J. S., 90, 311
Zigler, E., 85, 315
Zimmerman, W., 82, 313

Subject Index

ABAB design, 203
Ability test, 188
Abstract goals, 62, 81
Accountability, 41, 68
Achievement tests, 94
Acquiescence, 129
ADEs, *see* Avoidable drug errors
Advocacy, evaluation *vs.*, 102–103
AEA, *see* American Evaluation Association
Afghanistan Agricultural Extension Project II (AAEP II), 271, 272–273
After-school program, 157, 167
Agency records, 94
Agreements, evaluator and, 31
Alligators, 54
Alpha change, in pretest–posttest design, 190
American Evaluation Association (AEA), 15, 103, 198, 287
Analysis of covariance (ANCOVA), 206, 221
Analysis of variance (ANOVA), 206, 221
Answers, recording, 164
Antabuse, 56
Antidepressant medication, 66
Antipsychotic medication, 57
Appendices, of reports, 269
Appreciative inquiry (AI), 168, 278
Archives, *see* Records
Artifacts, 79, 156
Assumptions, in cost analyses, 253
Athletic program, 49
Attempts to compensate the control group, 230
Attitude
  of evaluators, 44, 285
  measurement research, 129
Attrition, 186, 228
Auditing, 20
Avoidable drug errors (ADEs), 283

Behavioral therapists, 201, 203

Behavior analysts, 201
Benchmarks, 192
Benefits
  compared to costs, 242
  psychological, 251
Beta changes, in pretest–posttest design, 190
Black box evaluation, 27
Bloodletting, 182
Budget, 250
  evaluation, 70
  example program, 241
Business, 20–21

Canadian Evaluation Society (CES), 15
Capacity-building approach, 28
Case studies, 130–131
Case study 1, 21–22
Case study 2, 44–45
Case study 3, 71–73
Case study 4, 98–99
Case study 5, 117–119
Case study 6, 138–140
Case study 7, 151–152
Case study 8, 171–173
Case study 9, 194–196
Case study 10, 216–217
Case study 11, 233–235
Case study 12, 253–255
Case study 13, 272–273
Case study 14, 288–290
CES, *see* Canadian Evaluation Society
Change
  kinds of, 190–191
  scores, 182–183
Checklists, 93–94, 287
Communication plan, 259–261
Community, 171–173
Community-level indexes, 80
Community mental health, 50, 57
Community setting, 126
Community surveys, 129–130
Comparison groups, 206, 207
  selecting appropriate, 207–209
Confidence interval, 88, 231–232, 266–267
Confidentiality, 105–106
Conflict
  among stakeholders, 277
  with program staff, 106–107

Construct validity, 190–191
Contracts, 40
Control construct design, 212
Control group, 230
Cooperation, 160
Corporate welfare, 68
Cost(s), 11
  comparing outcomes to, 242–247
  of evaluation, 169–170
  program, 67, 239–242
Cost analyses, 238
  criticisms of, 251–253
  future costs and benefits, 247–249
  issues in, 247–251
  necessity of, 242
Cost–benefit analysis
  assumptions, 253
  cost-effectiveness and, 250–251
  illustration of, 243–245
Cost–effective measures, 91
Cost-effectiveness analysis, 245
  assumptions, 253
  cost-benefit and, 250–251
Cost-effective programs, 109, 245
Costs and benefits
  comparing, 243
  future, 247–249
Cost-utility analysis, 246
Counseling programs, 179
Crime records, 89

Data analysis, 231–232
Data collection, 33, 39
Data Driven Solutions (DDS), 98, 150
Data-gathering procedures, 266
Data sources, for evaluation, 77
  artifacts, 79
  choice of, 84–85
  community indexes, 79–80
  evaluation staff, 82–83
  expert observers, 81
  program participants, 78–79
  program records, 80–81
  program staff, 80
  trained observers, 82
Demographic information, 265
Diffusion or imitation of the program, 230
Direct costs, 240

Direct observations, 174
Disulfiram, 228
Documents, 161
Drafts of reports, 263–264

Economic analyses, 251
Education, 6, 20, 55
Effectiveness, criteria of, 85
Effect size, 180, 181, 229, 232, 244, 268, 281
Efficiency, 107–108
Elimination of program, 42
Empowerment evaluation, 28, 60
Energy conservation programs, 52
Ethical challenges, for evaluators, 117
Ethical standards, 103–104
Ethics, 102
  Institutional Review Board, 115–117
  interests of different stakeholders, 107–109
  negative side effects of evaluation procedures, 112–115
  practice of evaluation, 103–104
  role conflicts facing evaluators, 106–107
  treatment of people, 104–106
  validity of evaluations, 109–112
Evaluability, of program, 35–36
Evaluation, *see also* Program evaluation
  advocacy *vs.*, 102–103
  avoid negative side effects, 112–115
  Community Mental Health Center, 102
  credibility of, 38
  criteria, 60–62
  ethics in, 104–106
  falsely negative, 113
  falsely positive, 112
  focus of, 33
  impact on programs, 44
  improvement-focused approach, 69–70
  issues in, 11
  for knowledge, 14
  objectives-based program, 65–66, 69
  program planning and, 51
  reason for, 34–35
  resources availability, 35
  service and, 18
  standards for, 103–104
  time for completion, 34–35
  validity of, 109–112
  variables relevant to, 86
Evaluation attitude, 285–287

Evaluation budget, 70
Evaluation models, 26–28
  characteristics of, 29
  improvement-focused approach, 30–31
  information needs, 32–36
  planning evaluation, 36–40
  program identification, 31–32
Evaluation report, 258, *see also* Personal presentations
  attractiveness, 270–271
  communication plan, development of, 259–261
  context, 265
  draft, 263–264
  findings, 266–269
  misinterpretations, 267–269
  outline and summary, 264–265
  purposes of, 264
  recommendations, 269–270
Evaluator profile 1 (Birkby), 23
Evaluator profile 2 (Langhout), 46
Evaluator profile 3 (Beacham), 73–74
Evaluator profile 4 (Stacy), 100
Evaluator profile 5 (Wohlfarth), 120
Evaluator profile 6 (Sheppard), 140
Evaluator profile 7 (Dever), 153
Evaluator profile 8 (Kniffley), 196
Evaluator profile 9 (Ferrari), 196–197
Evaluator profile 10 (Chapman), 217–218
Evaluator profile 11 (Dickie), 235–236
Evaluator profile 12 (Pennington), 256
Evaluator profile 13 (Castelin), 273–274
Evaluator profile 14 (Deck), 290–291
Evaluator's opportunity, 158
Evaluators, roles of, 14–21
"Executive Summaries", 264
Experimental designs, 220–222
Experiments
  benefits of, 220
  cost of not doing, 224
  objections to, 222
  in program evaluation, 220–222
Expert observers, 81
External evaluator, 15–16, 17

Facebook, 84
Feedback, program evaluation, 13, 224, 227
Findings
  evaluation reports, 266–269
  mixed, 279–280

personal presentations of, 261–264
Fitness programs, 179
Fixed costs, 239
Focus groups, 93, 132–133, 134, 162
Food stamps, 136
Formal written evaluation reports, *see* Evaluation report
Formative evaluation, 6, 39, 259, 265–266

Gamma changes, in pretest–posttest design, 190
Goals
abstract, 62, 81
implementation, 63–65
intermediate, 65
outcome, 65
seldom-stated, 66
types of, 63
Government, 12
Government Accountability Office, 19
Government Performance and Results Act (GPRA), 177
GPRA, *see* Government Performance and Results Act
Grants, 35
Graphs, in reports, 266–267
Group interviews, 93

Head Start, 60, 227
Health and Psychological Instruments (HAPI), 94
Health care programs, 136
Hidden agendas, 68
Hidden cost, 240
Hindsight bias, 191, 192, 262
History, 185, 230
Human resources, 21
Human service managers, 144
Hypothetical evaluation, 157, 259
Hypothetical mean scores, 215

Ignatian Spirituality Project (ISP), 194, 197
Imagine learning, 187
Impact model, 35–37, 53–54, 137, 174, 192, 262, 266, 278, 282
Impacts, of program, 10–11
Implausible program theories, 54
Implementation, 4, 10, 30, 45, 56, 57, 59, 60, 71–72, 109, 111, 118, 136, 139, 151–152, 157, 158, 169, 171, 195, 216, 233, 244, 254, 272, 278, 281, 288, 291
fidelity of, 106
goals, 63–65
Impressions, 165
Improvement, 178, 182, 195
level of, 58
needed to justify program, 228

program, 280, 283
Improvement focus, 6, 30–31, 69, 99, 119
Incidence, 125
Incremental costs, 239
Indirect cost, 240
Industry, 17
Ineffective programs, 104–105
Information
misuse, 42–43
needs, 259
sources of, 39, 81
Information gathering, 86
cost-effective measures, 91
measures detecting change, 89–91
nonreactive measures, 86
reliable measures, 86–88
valid measures, 88–89
Information systems, 10
problems in, 148–150
program records and, 147–148
summarizing information, methods of, 146–147
Informed consent, 105
Innovation, 42
using evaluations, 281–282
Institutional Review Board (IRB), 115–117
Instrumentation, 189, 216
Intangible costs, 244–245
Integrating theory, 59–60
Intermediate goals, 65, 137, 138
Internal evaluator, 16–17, 194, 285
Internal validity, 184–190, 199, 201, 212–216
Internet, 41, 63–65, 128
Interpersonal relations, 285
Interpretations, 165–166, 185, 189, 192
Interrupted time series, 202, 203, 213–214
Interviews, 92–93, 110, 161
asking questions, 162–163
developing rapport, 162
probing, 163–164
recording answers, 164
ISP, *see* Ignatian Spirituality Project

Job-training program, 10, 85, 245

Key Evaluation Checklist, 287
Key informants, 131–132

Learning culture, 282–285, 287
"Letter of the law", 156

Literature review, 36–38
Local history, 185, 230
Lowess line, 204, 205

Management information system, 144–145, 276
Marketing research, 132
Maturation, 184–185, 221, 225
Measurement, 77
Measures, multiple, 85, 86, 88
Medical care, 20
Medical residency program, 108
Meetings, 259–260
Merits, 30–31
MISSION, 8
Mixed design, 206
Mixed findings, 279–280
Models of evaluation, *see* Evaluation models
Monitoring, 4, 10, 14, 88, 137, 143–145, 153, 158, 225, 231, 287
Multiple measures, 85, 88
Multiple regression, 210, 211

National Science Foundation (NSF), 177
Natural comparison group, 208
Need, 3, 11, 71
  assessment of, 56, 123, 125–134
  community setting, 126–127
  concept of, 126
  context of, 135–136
  definitions of, 123–125
  failure to examine, 135–136
  ignorance of, 136–137
  implementation plans, 136
  incidence, 125
  key informants, 131–132
  long-term, 12
  meeting, 3–4
  planning process, 126
  potential, 12
  problems of, 134–137
  of program participants, 9
  in program planning, 137–138
  self-report survey, 126–127
  services analysis, 130–131
  short-term, 11–12
  social indicators of, 127–129
  unmet, 3–4, 9, 56
Negative evaluation, 224, 228, 238
Negative side effects, 66, 69, 72, 112–115, 222, 226, 280

New Math, 57
Nonequivalent control group designs, 205–207
Nonparticipant observation, 159, 160
Nonreactive measurement, 86
Nonrecurring cost, 240
No-treatment group, 207
NSF, *see* National Science Foundation

Objective behavior, 87, 88
Objectives-based evaluation, 27, 66, 69
Observations, unrestricted, 165
Observer, central importance of, 158–159
Obvious cost, 240
One-group designs
  construct validity, 190–191
  descriptive designs, 178–184
  evaluation designs, 176–177
  internal validity threats, 184–190
  overinterpreting results, 191–192
  usefulness of, 192–194
Open-ended questions, 96, 162
Open forums, 132, 133–134
Opportunity cost, 240–242
Oral presentations, 261, 262
  of reports, 263, 265, 269
Outcome, 2, 3, 4, 6–7, 11, 30, 36, 45, 50, 51, 58, 68–69, 80, 87, 106, 110, 137, 143, 157, 169, 192, 221, 246, 266, 273, 277, 284
  compared to costs, 242–247
  desired, 60
  final, 51–52
  goals, 65
  intermediate, 51
  levels of, 58
  over time periods, 202–203
Outcome evaluation, 231
  experimental, 220
  one group, *see* One-group designs
  quasi-experimental, 199
  validity of, 199
Output, 9, 201
Overhead, 240

Parents Too Soon, 59
Partial correlation, 182–184, 193
Participant observation, 160
PERMA theory, 71–73
Personal presentations
  audience for, 263
  content of, 262–263

distributing drafts of reports, 263–264
need for, 261
Physical fitness program, 61
Physical trace, 161
Planning process, evaluator, 126
Plan reporting meetings, 259, 260
Plans implementation, 136
Political campaign, 157–158, 167
Posttest-only, 177–178, 186
Power, statistical, 281
Press releases, 271
Pretest–posttest design, 177, 185, 186, 189, 192, 201, 205–206, 215
construct validity in, 190–191
use of, 193
Pretests, 221
instructions and, 97–98
Prevalence, 125
Probing, 163–164
Program cancellation, 42
Program description, 31
Program director, 32
Program evaluation, *see also* Evaluation
confused with, 18–19
definition, 1, 2
extensiveness of, 12–13
methods and objectives, 19
misunderstandings of, 40–41
monitoring function of, 4
planning, 20
purpose, 13–14
types of, 9–11
Program managers, constraints on, 277
Program participants, 27, 32, 39, 41, 78–79, 146–147
assess needs of, 9
description of, 265
written surveys and interviews with, 91–93
Program planning, 9, 51, 137–138
Program records, 80–81
Program resources, 43
Program sponsors, 32–33
Program theories
development of, 50–54
evaluation questions, 55–60
goals and objectives, 62–70
illustration, 53
implausible, 54
implementation, 56–58

Progress reports, 262, 271
Proposal, 40
Prozac, 66
Pseudo-participant, 160
Psychological benefits, 251
Public conflicts, 107
Public education, 68

Qualitative data, 156
Qualitative evaluation, 155–156
  carrying out, 164–167
  coordinating qualitative and quantitative, 167–170
  evaluation settings, 156–158
  gathering qualitative information, 158–164
  philosophical assumptions, 170
Qualitative evaluators, 158–159
Qualitative information, 158–164
Qualitative interviewing, 161–162
Quality assurance, 14, 17, 153
Quantitative evaluations, 39–40, *see also* Experimental designs; Pretest–posttest design
Quasi-experimental designs, 220
Questions
  evaluation, 169
  open-ended, 96, 162

Randomization, 221–222, 229–231
Rapport, 162
Ratings, 156
Reactivity, 86, 188–189
Recommendations, in reports, 269–270
Recording answers, 164
Records, 80–81, 94, 147–148, 161, 164
Recurring cost, 240
Recycling program, 60–61
Regression, 208
Regression discontinuity design, 209–211
Regression toward the mean, 186–187, 208, 225
Rehabilitation programs, 14
Reliability, 86–88
Remedial programs, 188
Report outline, 264–265
Reports, 166
Representative sample, 38–39, 129
Requests for proposals, 25
Research design, 199
  appropriate, 111
  choice of, 39
  limitations on, 10

Resentful demoralization, 230
Resources availability, 35
Rivalry with experimental groups, 230

Sample
  representative, 268
  size of, 228, 229
Sanders College, 259, 263
SAT, 89
Scatterplot, 205, 210
Selection, 185–186
Selection-by-maturation interaction, 189
Selective control design, 214–216
Self-esteem, 104–105
Self-report survey, 126, 190, 266
Sensitivity, of measures, 89–91
Service delivery, assistance in, 108–109
Services analysis, 130–131
Side effects, 5–6, 69–70, 86
  negative, 6, 66, 69, 80, 104, 112–115, 222, 226
  positive, 6, 21–22, 23, 45, 72, 99, 195, 234, 272, 289
Significant others, 82
Single-group designs, 192–193, 197
Single-subject research, 201
Smoothing a graph, 205
“Snowball sample”, 131
“Social benefits”, 244
Social indicators, of need, 127–129
Social justice, 68, 107
Social marketing techniques, 137
Social media, 84–85
Social security, 177
Social service programs, 55, 104
Social support, 52
Social welfare programs, 50
Staff
  evaluation, 82–83
  perceptions of, 43
  of programs, 80
Stakeholders, 4–5, 25, 31, 55, 67–68, 71, 86, 104, 221, 272, 278
  criteria credible to, 71
  information needs, 259
  interests of different, 107–109
  meeting with, 32
  plan reporting meetings, 259–260
  of publicly funded programs, 106
  in selecting evaluation criteria, 61–62
  value conflicts among, 277

working with, 227, 282–283
Standard deviations, 180, 229, 268–269
Statistical analyses, 39–40, 115, 206, 228
Statistical hypothesis testing, 231
Statistical power, 231, 266, 281
Statistical process control, 200
Stem-and-leaf plot, 267
Students Against Drunk Driving (SADD) evaluation, 230
Subjectivity, 167
Success case method, 67–168
Summative evaluation, 6, 13–14, 17, 33, 266, 283
Sunk cost, 239
Surveys, 39, 86, 87
format of, 96–97
instructions and pretests, 97–98
for needs assessment, 129–130
participant, 66
preparing, 96–98
with program participants, 91
written, 91–92

Telephone interviews, 93
Tentative indicators, 284–285
Tests, 94, 188–189
Theory, 49, 71–75
choosing evaluation criteria, 60–62
evaluation budget, 70
identifying goals and objectives, 62–70
program, 50–54
questions flow, 55–60
Theory-based evaluation, 27–28, 59
Time series, 200
effect of HIV test, 201
interrupted, 202, 203
interrupted with switching replications, 213, 214
nonequivalent control groups and, 212–214
Traces, 161
Trained observers, 82
t-test, 221, 231
Twitter, 84
Type I error, 112, 191, 192, 281
Type II error, 112, 113, 269, 281

Units of analysis, 247
Unrestricted observations, 165
Utilization
of evaluations, 276–291
obstacles to, 276–278

Validity, 87, 89
  of evaluations, 109–112
Valid measures, 88–89
Value, 3
Variable costs, 239
Variables
  relevant, 86
Vitamin C, 66, 124

Welfare policies, 5
Worth, 3, 26, 30
Written proposal, 40
Written surveys, 91, 92