PROGRAM EVALUATION

5E

ALTERNATIVE APPROACHES AND PRACTICAL GUIDELINES

JODY L. FITZPATRICK • JAMES R. SANDERS
BLAINE R. WORTHEN • LORI A. WINGATE

Program Evaluation

Alternative Approaches and Practical Guidelines

FIFTH EDITION

Jody L. Fitzpatrick
University of Colorado Denver

James R. Sanders
Western Michigan University

Blaine R. Worthen
Utah State University

Lori A. Wingate
Western Michigan University

Content Management: Bridget Daly
Content Production: Yagnesh Jani
Product Management: Drew Bennett
Product Marketing: Krista Clark
Rights and Permissions: Jenell Forschler

Please contact https://support.pearson.com/getsupport/s/ with any queries on this content

Cover Image by Shunli zhao/Moment/Getty Images

Library of Congress Cataloging-in-Publication Data

Names: Fitzpatrick, Jody L., author. | Sanders, James R., author. | Worthen, Blaine R., author. | Wingate, Lori A., author.
Title: Program evaluation : alternative approaches and practical guidelines / Jody L. Fitzpatrick, James R. Sanders, Blaine R. Worthen, Lori A. Wingate.
Description: Fifth edition. | Boston : Pearson, 2022. | Includes bibliographical references and indexes.
Identifiers: LCCN 2021058174 | ISBN 9780137547586 (paperback)
Subjects: LCSH: Educational evaluation—United States. | Evaluation research (Social action programs)—United States. | Evaluation—Study and teaching—United States.
Classification: LCC LB2822.75 .W67 2022 | DDC 379.1/58—dc23/eng/20220124
LC record available at https://lccn.loc.gov/2021058174

ScoutAutomatedPrintCode

Pearson

Rental

ISBN 10: 0-137-54758-7
ISBN 13: 978-0-137-54758-6

Pearson's Commitment to Diversity, Equity, and Inclusion

Pearson is dedicated to creating bias-free content that reflects the diversity, depth, and breadth of all learners' lived experiences.

We embrace the many dimensions of diversity, including but not limited to race, ethnicity, gender, sex, sexual orientation, socioeconomic status, ability, age, and religious or political beliefs.

Education is a powerful force for equity and change in our world. It has the potential to deliver opportunities that improve lives and enable economic mobility. As we work with authors to create content for every product and service, we acknowledge our responsibility to demonstrate inclusivity and incorporate diverse scholarship so that everyone can achieve their potential through learning. As the world's leading learning company, we have a duty to help drive change and live up to our purpose to help more people create a better life for themselves and to create a better world.

Our ambition is to purposefully contribute to a world where:

- Everyone has an equitable and lifelong opportunity to succeed through learning.
- Our educational content accurately reflects the histories and lived experiences of the learners we serve.
- Our educational products and services are inclusive and represent the rich diversity of learners.
- Our educational content prompts deeper discussions with students and motivates them to expand their own learning (and worldview).

Accessibility

We are also committed to providing products that are fully accessible to all learners. As per Pearson's guidelines for accessible educational Web media, we test and retest the capabilities of our products against the highest standards for every release, following the WCAG guidelines in developing new products for copyright year 2022 and beyond.

You can learn more about Pearson's commitment to accessibility at **https://www.pearson.com/us/accessibility.html**

Contact Us

While we work hard to present unbiased, fully accessible content, we want to hear from you about any concerns or needs with this Pearson product so that we can investigate and address them.

Please contact us with concerns about any potential bias at **https://www.pearson.com/report-bias.html**

For accessibility-related issues, such as using assistive technology with Pearson products, alternative text requests, or accessibility documentation, email the Pearson Disability Support team at **disability.support@pearson.com**

Pearson

About the Authors

Jody Fitzpatrick is an emeritus faculty member in public administration at the University of Colorado Denver, where she taught research methods and evaluation. She has conducted evaluations in many schools and human service settings and written extensively about successful evaluation practice. She served on the board and as president of the American Evaluation Association (AEA) and on the editorial boards of the *American Journal of Evaluation* and *New Directions for Evaluation*. She also chaired AEA's Teaching of Evaluation topical interest group and won a teaching award at the University of Colorado Denver. In her book, *Evaluation in Action: Interviews with Expert Evaluators*, she used interviews with expert evaluators to talk about the decisions that evaluators face as they plan and conduct evaluations and the factors that influence their choices. She evaluated the changing roles of counselors in middle schools and high schools and a program to help immigrant middle-school girls to achieve and stay in school. Her international work includes research on evaluation in Spain and elsewhere in Europe, and she has spoken on evaluation issues to policymakers and evaluators in France, Spain, Denmark, Mexico, and Chile.

James Sanders is professor emeritus of educational studies and former associate director of The Evaluation Center at Western Michigan University, where he taught, published, consulted, and conducted evaluations. A graduate of Bucknell University and the University of Colorado, he served on the board and as president of the American Evaluation Association (AEA). He chaired the steering committee that created the Evaluation Network, a predecessor to AEA. His publications include books on school, student, and program evaluation. He has worked extensively with schools, foundations, and government and nonprofit agencies to develop their evaluation practices. As chair of the Joint Committee on Standards for Educational Evaluation, he led the development of the second edition of the Program Evaluation Standards. He was also involved in developing the concepts of applied performance testing for student assessments, cluster evaluation for program evaluations by foundations and government agencies, and mainstreaming evaluation for organizational development. His international work in evaluation concentrated in Canada, Europe, and Latin America. He received distinguished service awards from Western Michigan University, where he helped establish a Ph.D. program in evaluation, and from the Michigan Association for Evaluation.

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

Lori Wingate is the executive director of The Evaluation Center at Western Michigan University (WMU), where she has worked since 1997. She has a Ph.D. in interdisciplinary evaluation from WMU and has been working in the field of program evaluation since 1993. From 2009 to 2019, she directed EvaluATE, the evaluation hub for the National Science Foundation's Advanced Technological Education program. From 2011 to 2019, she served as a subject matter expert in evaluation for the U.S. Centers for Disease Control and Prevention (CDC). She has led more than 75 webinars and workshops on evaluation in various contexts, including CDC University, American Evaluation Association Summer Evaluation Institute, and EvaluATE. She leads the Evaluation Checklist Project at WMU and has developed numerous resource materials to support evaluation practice, including checklists, templates, and guides. She has written book chapters on evaluating humanitarian response to emergencies, evaluation checklists, and metaevaluation. She was the book review section editor for the *American Journal of Evaluation* from 2005 to 2011 and has led a range of evaluation projects in the areas of STEM education, public health, and higher education.

About This Book

The first edition of this book was published in 1987. At that time, we had a vision for a book that would provide a solid foundation for anyone interested in conducting or using evaluation to promote the public good. Although the evaluation field has grown and changed immensely over the past 35 years, we have not wavered from that vision. With each new edition, we have updated the content of the previous edition and added new material that we believe added value to what had been published before. This continues to be the case for this fifth edition.

We are proud to announce the addition of a new coauthor on our team. In addition to Drs. Worthen and Sanders, who come with an education perspective, and Dr. Fitzpatrick, who provides a public administration perspective, new coauthor Dr. Lori Wingate adds a multidisciplinary perspective to the team and a strong track record of providing practical guidance on evaluation to diverse audiences. We all share a common vision of the promise of evaluation for the public good.

We hope this book will inspire you to think in new ways about programs, policy, and organizational change. If you are already an evaluator, this book will provide you with new perspectives and tools for your practice. If you are new to evaluation, it will make you a more informed consumer of evaluation studies. If you are an aspiring evaluator, it will help you form a strong foundation for your future professional development.

The book has four parts. Part 1 introduces key concepts in evaluation, historical developments and current trends in the field, and core tenets of professional program evaluation. In Part 2, we present several different approaches to evaluation, often called models or theories. Evaluators must know methodology, but they also must know about the theoretical foundations of evaluation. Evaluation approaches serve as useful heuristics for creating evaluation plans that are appropriate for specific evaluation contexts, programs, clients, and stakeholders. In Parts 3 and 4, we describe how to plan and carry out program evaluations. Part 3 focuses on the planning stage: learning about the program, conversing with stakeholders to understand the purposes and expected uses of the study, identifying evaluation questions, and developing a management plan to guide the study. In Part 4, we discuss the methodological choices that evaluators make about selecting and developing designs, sampling, collecting and analyzing data, interpreting results, and communicating about the evaluation. The chapters in each of these sections are sequential, representing the order in which evaluators usually make decisions in an evaluation.

Each chapter begins with orienting questions to introduce its main topics and wraps up with a list of the key points. To extend learning, each chapter concludes with discussion questions, application exercises, and a list of suggested resources to learn more.

New to This Edition

- A new chapter (Chapter 3) on core tenets of program evaluation, covering the profession's standards and principles and the competencies it requires.
- In Chapters 5 through 8, embedded descriptions of real-world evaluations that illustrate applications of the approaches described in these chapters.
- In each chapter's Suggested Resources section (formerly Suggested Readings), a greater variety of recommended materials (e.g., websites, videos, checklists, etc.), and annotations highlighting the most useful features of each resource.
- More tables, graphics, and headings throughout the chapters to help with navigation and identification of key takeaways.
- Reorganization of some of the content on evaluation approaches (Chapters 5 through 8) to distinguish between contemporary approaches and their historical foundations.

Key Content Updates

Changes in the fifth edition include the following:

- Chapter 1: Reorganized and streamlined content.
- Chapter 2: Added information about notable developments in the field since the last edition of this book was published in 2011. Added information about contributions of early Black evaluators and education researchers.
- Chapter 3: Created this new chapter on the core tenets of program evaluation, covering the profession's standards and principles and the competencies expected of evaluators.
- Chapter 4: Added graphics and tables to reinforce main concepts (no major changes to content).
- Chapter 5: Reframed the collection of evaluation approaches described in this chapter as "judgment-oriented approaches."
- Chapter 6: Separated the discussion of historical foundations of program-oriented approaches from the discussion of contemporary approaches; added a new section on the Kirkpatrick model for evaluation.
- Chapter 7: Moved information about evaluability assessment to Chapter 11 to accompany other information about clarifying the evaluation request.
- Chapter 8: Separated the discussion of historical foundations of participation-oriented approaches from the discussion of contemporary approaches; reorganized and expanded content on participation-oriented evaluation approaches to include collaborative evaluation, transformative participatory evaluation, and research-based principles of collaboration in evaluation; added a table to underscore the core conceptual differences among these approaches.

- Chapter 9 (Chapter 10 in fourth edition): No major changes to this chapter that compares the approaches reviewed in Chapters 5 through 8.
- Chapter 10: Expanded the previous edition's content on culture, including cultural competence and culturally responsive evaluation.
- Chapter 11: Moved content on evaluability assessment here from an earlier chapter in the fourth edition; moved previous edition's content on evaluation use to Chapter 18.
- Chapter 12: Expanded the information about logic models, with additional guidance and examples.
- Chapter 13: No major changes.
- Chapter 14: Streamlined information about designs, data collection sources, and methods to minimize overlap with later chapters; updated the evaluation budget example.
- Chapter 15: Reframed discussion of study design options in terms of experimental, quasi-experimental, and nonexperimental designs; and expanded the information about sampling.
- Chapter 16: Added information about program artifacts as data sources; specific mixed-methods designs; survey questions; tradeoffs among paper, electronic, telephone, and in-person surveys; and innovations in data collection, analysis, and interpretation.
- Chapter 17: Reorganized and substantially updated information about reporting to reflect advances in this aspect of evaluation practice.
- Chapter 18: Substantially revised chapter to focus on the future of the evaluation profession, evaluation practice, evaluation education, and research on evaluation.

Pedagogical Features

Each chapter begins with **Orienting Questions** to introduce its main topics and wraps up with a corresponding list of **Key Points**.

To extend learning, each chapter includes **Discussion Questions**, **Application Exercises**, and **Suggested Resources**. The discussion questions are intended to prompt reflection and discussion related to each chapter's content. The application exercises provide a structure for students to engage with the material in ways that simulate the decisions and critical thinking involved in real-world evaluation practice. The annotated suggested resources offer students many options for learning more about key topics covered in each chapter.

In Chapters 5 through 8, **Case Descriptions** of real-world evaluations illustrate applications of the evaluation approaches described in these chapters.

Acknowledgments

We would like to thank our colleagues in evaluation for continuing to make this such an exciting and dynamic field. With each revision of this book, we are reminded of the progress being made in evaluation and in our colleagues' wonderful insights about evaluation theory and practice. We all are grateful to our families

for the interest and pride they have shown in our work and the patience and love they have demonstrated as we have devoted our time to it.

We also thank our copyeditor Carolyn Williams-Noren for her skilled assistance in enhancing the clarity and organization of our writing.

Our newest coauthor, Lori Wingate, extends special thanks to her Western Michigan University (WMU) colleagues. Chris Coryn and Michael Harnar provided valuable feedback on draft chapters about design and participation-oriented evaluation. Kelly Robertson and Lyssa Wilson Becho provided useful insights about culturally responsive evaluation, mixed-method designs, and how evaluators really use evaluation approaches in practice. They also indulged her in countless conversations about various evaluation concepts and practices that informed the revision of this book. Lori especially thanks her family—Scott, August, and Stuart—for their love, patience, and understanding.

Brief Contents

Contents

Part 1

Introduction to Evaluation

This opening section provides background information that will help you understand the chapters that follow. We have included references throughout the chapter that point to a wealth of material about the foundations of the evaluation discipline.

In Chapter 1, we provide a detailed definition and explanation of what program evaluation is. We introduce several core evaluation concepts. We distinguish evaluation from research and other forms of inquiry and give several examples of program evaluations.

Chapter 2 summarizes the origins of evaluation in the United States as a distinct form of inquiry and professional practice. We review the field's development, noting significant milestones in the growth of evaluation as a force for improving public, nonprofit, and corporate programs.

In Chapter 3, we review the evaluation profession's standards and principles and the competencies required of evaluators. While we emphasize evaluation practice in the United States, we also point to sets of guidelines and standards that pertain to evaluation in several other parts of the world.

These chapters lay the foundation for Part 2, where we introduce several evaluation approaches to expand your understanding of the breadth of choices that evaluators and stakeholders may make.

This page intentionally left blank

1

Introduction to Program Evaluation

Orienting Questions

1. What is program evaluation?
2. What are the main purposes and types of evaluation?
3. How are research and evaluation alike and different?
4. Other activities—like monitoring, policy analysis, and quality assurance—seem a lot like program evaluation. How are they related?
5. What can evaluation do for organizations, communities, and society? What are its limitations?

Throughout the world, institutions and organizations develop and implement programs to tackle problems and enhance conditions in diverse areas. Governmental agencies, nongovernmental and nonprofit organizations, and business and industry groups are grappling with complex issues, including how to

- reduce income inequality and structural racism
- increase racial and gender equity
- decrease food insecurity
- enhance access to higher education and prepare students for the workforce
- help veterans transition back to civilian life
- combat disease and mental illness
- reduce crime and improve the justice system
- mitigate the effects of climate change.

And that just scratches the surface!

Some programs are modest in size, such as a local nonprofit's program to provide economically disadvantaged elementary school students with winter coats and boots. Some programs are ambitious and far-reaching, such as a federal agency's efforts to increase the diversity of the national STEM (science, technology, engineering, and math) workforce or reduce the incidence of chronic disease. All kinds of programs can benefit from systematic evaluation to ensure they are implemented effectively and equitably, identify ways they can be improved, determine the magnitude and importance of their outcomes, or assess their cost-effectiveness or sustainability.

The need for sound evaluation that produces credible and useful information becomes increasingly acute as resources to support programs grow scarce. Policymakers and program managers face tough choices about where and how to allocate funding. To make such choices intelligently, decision-makers need good information about the relative effectiveness of programs. Which programs are working well? Which are failing? What are the programs' relative costs and benefits?

Similarly, program developers and managers need to know how well different program components are working. What can be done to improve aspects of the program that are not working as well as they should? Have all aspects of the program been thought through carefully at the planning stage, or is more planning needed? Does the program work as expected, and how should it be adapted to respond to unexpected constraints or opportunities?

Answering such questions is the central task of program evaluation. The purpose of this book is to introduce you to program evaluation and orient you to the technical and practical aspects involved in evaluation. As a student of evaluation, you may aspire to become a professional evaluator. Or maybe you want to develop your evaluation literacy so you can be a more informed consumer of evaluation in your primary profession. We designed this book to

- equip you with a solid grounding in the foundations of program evaluation
- acquaint you with the history of the profession
- introduce you to a range of approaches and models that may be used to guide evaluation practice
- orient you to the methods and practices commonly employed in evaluation.

Most importantly, we hope this text helps you cultivate an evaluative mindset.

Definition of Program Evaluation

Program evaluation is a systematic process to determine the merit, worth, or significance of a coordinated set of activities designed to bring about change. Central to this definition are the terms *merit, worth*, and *significance:*

- "Merit is the excellence of an object as assessed by its intrinsic qualities or performance" (Yarbrough et al., 2011, p. 289).

- "Worth is the value of an object in relationship to needs or identified purposes" (Yarbrough et al., 2011, p. 293).
- "Significance refers to a program's potential influence, importance, and visibility" (Stufflebeam & Coryn, 2014, p. 708).

Something can have high merit, but low worth. For example, if an afterschool program had exceptional teachers and met all standards for out-of-school-time programming, it would have high merit. But suppose that program were located in an affluent school where most students already had afterschool care or scheduled activities such as clubs, sports, or music. In that case, it would not be valuable because it wouldn't be needed by those it was intended to serve. If the program offered especially innovative activities or brought together partners from diverse sectors that hadn't collaborated before, it might be judged to be highly significant.

To delve more deeply into the definition of program evaluation, we offer a more in-depth discussion of what we mean by *evaluation*, *program*, and *program evaluation*.

Evaluation

If you look up the word *evaluation* in a dictionary, you'll find a definition like this one from the Merriam-Webster online dictionary: "the determination of the value, nature, character, or quality of something or someone." This basic definition covers a wide range of human activity—from quick, informal judgments in daily life to formal, systematic assessments based on careful collection and analysis of evidence.

Evaluation is something you do in everyday life. You evaluate when you decide whether a new acquaintance is worthy of your continued attention. You evaluate when you decide which grocery items have the optimal combination of nutrition, flavor, and value. You evaluate when you judge the quality of a customer service experience.

Evaluation is also a responsibility of many professionals. For example, a school principal observes a teacher working in the classroom and forms judgments about that teacher's effectiveness. A program officer of a foundation visits a program for people with substance abuse disorder that the foundation supports and forms a judgment about the program's quality and effectiveness. A policymaker hears a speech about a new method for delivering healthcare to uninsured children and draws conclusions about whether it would work in their state. Many professionals make these types of evaluative judgments on a routine basis in their work. These judgments, however, are often based on relatively informal and unsystematic evaluations.

In contrast, formal, professional program evaluation involves being clear about what one is investigating, systematically gathering and analyzing data, and forming defensible conclusions in a transparent manner. "Formal evaluation," wrote Stake (2013) "is the conscious disciplining of judgment" (p. 189). As a human activity, evaluation exists on a continuum from informal to formal. In daily

life, it tends to be impressionistic and private. As a professional practice, it should be systematic and transparent. This book's focus is on professional evaluation.

The American Evaluation Association (AEA), the flagship organization of professional evaluators in the United States, defines evaluation as "a systematic process to determine merit, worth, value or significance" (AEA, 2014). This systematic approach sets formal, professional evaluation apart from the type of informal evaluation that is a fundamental, everyday human activity.

Programs and Similar Efforts

Pretty much anything can be evaluated. This book's focus is on the evaluation of programs. A *program* is an ongoing activity or set of related activities intended to bring about a change or improvement in a specified condition (or sometimes to prevent a condition from worsening). A *project* is similar but tends to be more time-bound, with a distinct endpoint. Likewise, *intervention* and *initiative* are more general terms that refer to organized efforts to bring about change or improvement. These labels are inconsequential for evaluation—what one person or organization calls a program, another may call a project or initiative.

Those who desire a more operational definition of *program* may appreciate the detailed definition provided by the Joint Committee on Standards for Education Evaluation (Yarbrough et al., 2011), which defines a program as

- a set of planned systematic activities
- using managed resources
- to achieve specified goals
- related to specific needs
- of specific, identified, participating human individuals or groups
- in specific contexts
- resulting in documentable outputs, outcomes, and impacts
- following assumed (explicit or implicit) systems of beliefs (diagnostic, causal, intervention, and implementation theories about how the program works)
- with specific, investigable costs and benefits. (p. xxiv)

Policy generally refers to a broader intention of a public organization or a branch of government. Organizations have policies; they guide practice related to recruiting and hiring employees, compensation, performance reviews, and more. Governmental bodies—legislatures, departments, executives, and others—also establish policies, which may take the form of laws or regulations. Sometimes, the line between a program and a policy is fuzzy. Like a program, a policy is designed to achieve some outcome or change. Unlike a program, a policy does not provide a service or activity. Instead, it provides guidelines, regulations, or the like to bring about change. However, programs may be established in response to a policy. Those who study public policy define it even more broadly as "the sum of government activities, whether pursued directly or through agents, as those activities have an influence on the lives of citizens" (Peters, 2018, p. 4).

Clear delineations and agreed-upon definitions of the terms *programs*, *projects*, *policies*, *interventions*, and *initiatives* do not exist. This book's focus is on the evaluation of coordinated efforts to bring about changes or improvements in specific conditions that affect humans and the well-being of the planet. We refer to such efforts as *programs* throughout this book.

This book does not address the evaluation of *products* or *personnel*, which are markedly different from programs, policies, projects, and the like. A product is a more concrete entity than a program. Programs sometimes generate products such as software, apps, websites, training manuals, and educational materials. Likewise, programs have personnel who design, manage, and deliver services. Therefore, evaluations of programs may include direct or indirect assessments of the effectiveness of products and personnel. However, stand-alone product and personnel evaluations involve different approaches and methods than do program evaluations.

Program + Evaluation = Program Evaluation

We have explained what we mean by *program* and *evaluation*. So it should not surprise you that *program evaluation* is a systematic process to determine a program's merit, worth, or significance. The Joint Committee on Standards for Educational Evaluation (Yarbrough et al., 2011) offered this detailed definition of program evaluation:

- the systematic investigation of the quality of programs, projects, subprograms, subprojects, and/or any of their components or elements, together or singly
- for the purpose of decision making, judgments, conclusions, findings, new knowledge, organizational development, and capacity building in response to the needs of identified stakeholders
- leading to improvements and/or accountability in the users' programs and systems
- ultimately contributing to organizational or social value. (p. xxv)

Building on the work of Scriven (1980), Fournier (1995) succinctly articulated the "general logic of evaluation," which helps illuminate the systematic nature of formal program evaluation highlighted above. This general logic unfolds in four steps:

1. Determine criteria of merit that define the desirable characteristics of the program being evaluated.
2. Establish standards of performance against which the program's quality will be judged.
3. Gather evidence of the program's performance.
4. Compare the program's performance with the established criteria and standards to reach a judgment about program quality.

As will become apparent in the rest of this chapter and this book overall, program evaluation is practiced in myriad ways. A program's context, the need for an evaluation, the values of program stakeholders, the background and expertise of evaluators, and many other factors come into play in shaping a given program evaluation.

Evaluation Stakeholders

We mentioned *stakeholders* a few times in the previous section. This term comes up frequently in evaluation. Stakeholders are individuals and groups who have an interest in and may be affected by the program being evaluated or the evaluation's results. Typical stakeholder groups include the program's funder, designer, managers, staff, and participants. Other groups more removed from the program may also be stakeholders, such as participants' family members or colleagues, organizations that provide similar services or serve similar populations, and groups that oppose the program or experience negative consequences because of it.

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

Main Purposes of Program Evaluation

Program evaluations usually serve one or more of the following purposes:

- *Developmental evaluation* takes place while a program is being developed to inform its design, testing, modification, or improvement.
- *Formative evaluation* occurs while a program is underway to identify ways it can be improved.
- *Summative evaluation* happens during program implementation, near the conclusion of program, or after it has ended. A summative evaluation's main purpose is to judge a program's merit, worth, or significance to inform decisions about whether to expand, continue, modify, or cancel it.

The distinctions between developmental, formative, and summative evaluation have to do with the kinds of decisions stakeholders expect to make based on the evaluation's results. With stakeholders, evaluators determine the relative emphasis on developmental, formative, or summative evaluation at the start of a study. Knowing the purpose of a study informs decisions about what types of data to collect, which stakeholders should be engaged, and who should receive the evaluation results. Table 1.1 compares developmental, formative, and summative evaluation in terms of the timing of an evaluation and the use of the evaluation results. Figure 1.1 illustrates these concepts using a cooking metaphor.

TABLE 1.1 Main Program Evaluation Purposes

Main purpose	When the evaluation is conducted	Use of the evaluation results
Developmental	While a program is being developed or during early implementation	To inform decisions about how to design or further develop a program
Formative	While a program is being implemented	To inform decisions about how to improve a program
Summative	While a program is being implemented, near the end of a program, or after a program concludes	To inform decisions about whether to continue, expand, contract, or discontinue a program

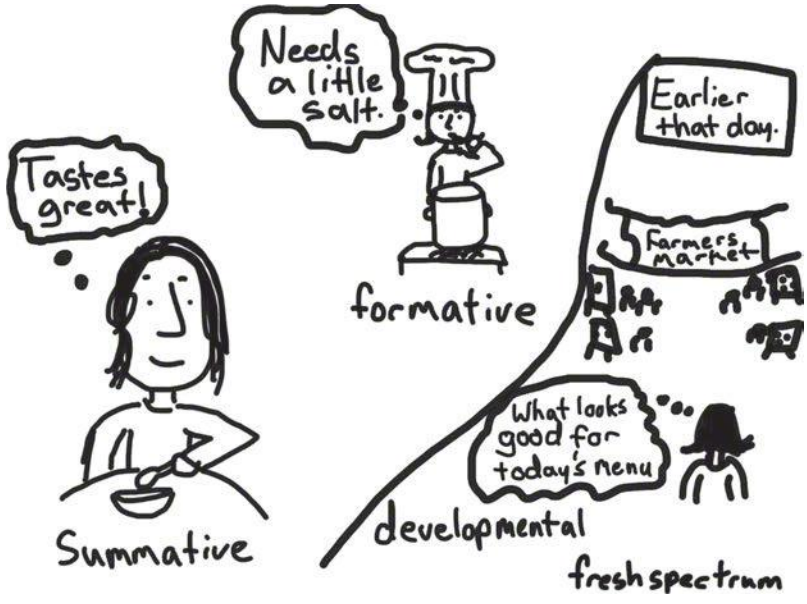

FIGURE 1.1 Illustration of differences between summative, formative, and developmental evaluation

Source: Lysy, C. (2020). Summative, formative, and developmental. *FreshSpectrum*. https://freshspectrum.com/summative-formative-and-developmental/

Main Types of Evaluation

Although this is not an exhaustive list, the main types of evaluation are *process evaluation*, *outcome evaluation*, and *cost analysis*. These labels are cues to the nature of the issues or questions a study will address.

Process Evaluation

Process evaluations focus on the content, outputs, or delivery of a program. Davidson (2005) described process evaluation as "taking a critical look at the quality or value of everything about the program (what it is and does) *except* outcomes and costs" (p. 56). Linnan and Steckler (2002) identified seven aspects of a program that might be assessed in a process evaluation, as shown in Table 1.2.

Because formative evaluation must occur while a program is being implemented, it is often confused with process evaluation. However, it is possible to conduct a summative evaluation of a program's processes. Such a study would deliver conclusions about the overall quality of a program's content and delivery to inform decisions about program continuation, expansion, or cancellation. A formative process evaluation would identify strengths and weaknesses of program implementation so that it could be improved while underway.

TABLE 1.2 Key Process Evaluation Components

Component	Definition
Context	Aspects of the larger social, political, and economic environment that may influence intervention implementation.
Reach	The proportion of intended target audience that participates in an intervention. If there are multiple interventions, then it is the proportion that participates in each intervention or component. It is often measured by attendance. Reach is a characteristic of the target audience.
Dose delivered	The number or amount of intended units of each intervention or each component delivered or provided. Dose delivered is a function of efforts of the intervention providers.
Dose received	The extent to which participants actively engage with, interact with, are receptive to, and/or use materials or recommended resources. Dose received is a characteristic of the target audience and it assesses the extent of engagement of participants with the intervention.
Fidelity	The extent to which the intervention was delivered as planned. It represents the quality and integrity of the intervention as conceived by the developers. Fidelity is a function of the intervention providers.
Implementation	A composite score that indicates the extent to which the intervention has been implemented and received by the intended audience.
Recruitment	Procedures used to approach and attract participants. Recruitment often occurs at the individual and organizational/community levels.

Source: Linnan, L., & Steckler, A. (2002). Process evaluation for public health interventions and research: An overview. In Linnan & Steckler (Eds.), *Process evaluation for public health interventions and research* (p. 12). Jossey-Bass.

Outcome Evaluation

Outcome evaluations measure changes among program participants, secondary impactees (such as participants' families or colleagues), organizations, communities, or larger segments of society. These outcomes may range from immediate program effects to longer-term results, both intended and unintended. *Outcomes* may include positive or negative changes in individual-level knowledge, attitudes, capabilities, or behaviors, or broader economic, health, social, or environmental conditions. *Impact* is a term used by some to refer to the ultimate long-term changes intended or realized by a program. Thus, an impact evaluation is a type of outcome evaluation. (Keep in mind that there is little consensus in the field about what counts as impact vs. an outcome; these terms are often used interchangeably.)

An outcome evaluation may serve formative or summative purposes. An evaluation designed to identify ways to improve program results would be a formative outcome evaluation. For example, teachers and trainers often use immediate measures of student learning to make changes in their curriculum or methods. They may decide to spend more time on some topics or activities to support students' achievement of specific learning goals. Or they may spend less time on areas in which students have already achieved competency. In contrast, policymakers making summative decisions are usually more concerned with the program's success at achieving higher-level outcomes. They may want to know about graduation rates or employment placement because they are more accountable for these outcomes.

Cost Analysis

Evaluations that examine program costs are less common than process and outcome evaluations. Analyzing program costs and benefits is a complex undertaking. Understandably, those who allocate resources to programs—whether private foundations, public agencies, or business enterprises—are very concerned with program costs. Cost studies, though somewhat rare, are important. There are two main types of cost studies: cost-effectiveness and cost-benefit analysis, which we describe in more detail in Chapter 15.

Cost-effectiveness studies compare the costs and outcomes of different programs designed to achieve the same or similar results. The programs must have the same costs so the outcomes can be compared, or the same outcome so costs can be compared (Alkin & Vo, 2018). In complex programmatic environments, such conditions are rare. Therefore, despite their attractiveness to decision makers who must decide how to utilize scarce resources, due to feasibility issues, cost-effectiveness studies are not common.

Cost-benefit studies involve identifying all costs and benefits associated with two or more programs and translating any nonmonetary costs and benefits into monetary units (e.g., dollars). Each program's costs are determined, and the benefits are identified and monetized. With such data, cost-benefit ratios can be calculated for

each program and compared. Not surprisingly, it can be challenging to translate all program benefits into dollar terms. Levin and McEwan (2001) cautioned that cost-benefit analyses are appropriate only "when the preponderance of benefits could be readily converted into pecuniary values or when those that cannot be converted tend to be unimportant or can be shown to be similar among the alternatives that are being considered" (p. 15).

Reconciling Types and Purposes of Evaluation

Labeling an evaluation as developmental, formative, or summative indicates how information from the evaluation is expected to be used. Saying an evaluation will be a process evaluation, outcome evaluation, or cost analysis identifies what aspects of a program will be investigated.

It is not unusual for people to mistakenly treat the term *formative* as synonymous with process evaluation, and *summative* as synonymous with outcome evaluation. However, Scriven (1996b), who coined the terms *formative* and *summative*, noted that "formative evaluations are not a species of process evaluation. Conversely, summative evaluation may be largely or entirely process evaluation" (p. 152). Table 1.3 illustrates how process, outcome, and cost evaluations may be either formative or summative. Developmental evaluation is not included in this table, because it occurs while a program is being formed—often before its processes, outcomes, or costs are established and able to be evaluated.

TABLE 1.3 Examples of How Process, Outcome, and Cost Evaluation May Be Either Formative or Summative

What the evaluation will focus on	Examples of how the evaluation results will be used	
	Formative evaluation	**Summative evaluation**
Program processes	To assess efficiency of program delivery to determine how it can be improved	To determine whether there is sufficient demand for a program to warrant its expansion
Program outcomes	To measure program effects on participants to identify whether the program should be modified to better serve all members of the intended audience	To determine whether the program is achieving sufficient results to justify continuation
Program costs	To compare costs and outcomes of different modalities of service delivery to determine which is optimal	To compare costs and outcomes of competing programs to determine which is best

Comparing Evaluation with Research

A question that often arises when individuals are first learning about evaluation is *How does evaluation differ from social or educational research?* There is no simple answer, because while there are differences, there are also commonalities. As Scriven (2016) noted, while research and evaluation are distinct, they are not dichotomous endeavors. There is no litmus test to determine whether an activity is evaluation or research. Below we describe how research and evaluation are alike and different in terms of the methods they employ, the main purposes they serve, their characteristics of practice, and their quality standards.

Methods

Evaluators draw heavily on qualitative and quantitative data collection and analysis methods developed in education, social science, and market research contexts. Examples include surveys, interviews, focus groups, observations, experiments, and analysis of secondary data such as test scores and census data. Because these methods are ubiquitous in both research and program evaluation, the two types of activities may seem indistinguishable from an outside perspective.

As program evaluation has developed and expanded into different sectors, practitioners have developed and adapted new methods for use in evaluation. Examples include appreciative inquiry, the most significant change technique, outcome harvesting, and goal attainment scaling. We describe these and other methods not typically used in research in Chapter 16.

Evaluators are eclectic in their methodological choices and almost always utilize traditional educational or social research methods such as surveys, tests, and observations to some degree. In contrast, researchers rarely use the methods developed specifically for program evaluation.

Purposes

Research and evaluation are typically undertaken for different purposes. The primary purposes of research are to add knowledge to a field and to contribute to the growth of theory. A good research study should advance knowledge and understanding of a topic. While the results of an evaluation study may contribute to knowledge development (Mark et al., 2000), that is a secondary concern in evaluation. Program evaluation's primary purpose is to provide useful information to stakeholders, often helping them to make a judgment or decision about the program's value. Some evaluators distinguish between research and evaluation with the saying, *Research tells you what's so; evaluation tells you so what.* Researchers may reach conclusions about whether a type of intervention has an effect (what's so) but typically do not ascribe importance or value to the effect (so what). Evaluations are usually conducted to reach conclusions about the importance, effectiveness, quality, or value of a specific program in a given context. In contrast, research is usually designed to produce results that can be generalized to other contexts.

A gray area in this distinction between the purposes of research and evaluation is *action research.* Action research is inquiry conducted collaboratively by professionals to improve their practice (Lewin, 1946). Such professionals might be social workers, teachers, accountants, or others who use research methods to investigate the effectiveness of their work. On the surface, action research may look similar to program evaluation, but there are key differences. Professionals engage in action research about their own work with the goal of improving their practice. Action research may also be used as a strategy to encourage professionals to work together to learn, examine, and research their own practices. Thus, action research produces information for improvement. The research is conducted by those delivering the program and, in addition to improving the element under study, has major goals concerning professional development and organizational change.

The primary purposes of program evaluation and research *tend* to be different, but they frequently overlap. For example, results of program evaluations can contribute to an evidence base about a particular type of intervention. Research results may inform decision-making about how best to address problems.

Practice

By *practice*, we mean how evaluators go about their work. As noted above, researchers and evaluators use many of the same methods for data collection and analysis. Notable ways in which the practices of evaluation and research differ have to do with who sets the agenda, who is involved, and how results are communicated.

In research, an individual researcher or research team usually sets the agenda for the inquiry. Researchers decide on research questions and hypotheses based on what they want to investigate and their judgments about what is needed to advance knowledge in their disciplines. In program evaluation, the questions to be answered do not originate with evaluators. Instead, they come from many sources, including the program's stakeholders. An evaluator might suggest questions but would never determine a study's focus without consulting stakeholders.

Good evaluation almost always involves stakeholders. Stakeholders are involved in evaluation for many reasons. Their engagement helps ensure that the evaluation addresses their information needs, builds organization capacity, and increases the likelihood that they will believe and use the results. In social and educational research, researchers may seek cooperation from individuals in the research context. However, the purpose is usually for facilitating the research, not setting the research agenda.

Research is conducted to advance knowledge within or across disciplines. Therefore, unless done for proprietary purposes, research is usually communicated through formal reports or white papers for broad dissemination or in peer-reviewed journal articles. Evaluators share evaluation results in more varied ways. These may include formal, public reports and journal articles, but more common are reports intended mainly for a program's stakeholders. These reports may take the form of informal verbal reports, memos, presentations, multiple short reports, or technical reports. Increasingly, both evaluation and research results are being conveyed via data visualizations and infographics for broad consumption.

Quality Standards

Research and evaluation differ in the standards used to judge their adequacy.

The traditional criteria for assessing research quality are validity, generalizability, reliability, and objectivity (Coryn, 2007). These criteria emerged from quantitative research. An alternative set of criteria for qualitative research includes credibility, transferability, dependability, and confirmability (Guba & Lincoln, 1989). These criteria for judging research quality focus on the properties of the research methods and findings.

The quality of program evaluations is judged based on other factors in addition to the validity or accuracy of data. Program evaluations are judged by their utility, feasibility, propriety, and accountability, as well as their accuracy (Yarbrough et al., 2011). The Joint Committee on Standards for Educational Evaluation defined standards for assessing these dimensions of evaluation quality. These standards help both evaluation users and evaluators to assess the quality of evaluations. These standards were originally developed for educational evaluations. However, the relevance of these standards in other contexts is demonstrated by their inclusion in the Centers for Disease Control and Prevention's evaluation framework (1999) and their adoption by professional evaluation organizations in Africa, Europe, and Latin America whose members work in diverse sectors. (We discuss these evaluation standards further in Chapter 3.)

In short, data validity is a necessary but not sufficient condition for a high-quality evaluation. Equally important is how an evaluation is carried out.

Table 1.4 highlights the key differences between research and evaluation in terms of their purpose, methods, practice, and quality standards.

TABLE 1.4 Key Differences Between Research and Evaluation

	Evaluation	Research
Main purposes	Make judgments about program merit, worth, or significance; provide information for decision making	Add to knowledge base in a field, develop and test theories
Methods	Eclectic	Typically defined by the discipline in which the research occurs. Traditional qualitative methods include observations, interviews, and focus groups. Traditional quantitative methods include surveys, experiments and quasi-experiments, and analysis of secondary data
Practice	• Questions determined by stakeholders' information needs • Significant involvement by stakeholders • Results communicated through a variety of methods	• Questions tend to be researcher-driven • Minimal involvement by stakeholders, if any • Results communicated mainly through peer-reviewed journal articles and public reports
Standards of quality	Utility, feasibility, propriety, accuracy, and accountability	Validity, reliability, objectivity, generalizability; credibility, transferability, dependability

Some evaluation students may feel frustrated with the lack of a single, clear dividing line between evaluation and research. Rather than deliberating over how to distinguish them, our advice is to make sure you know the basic tenets and tools of each and learn how to apply them appropriately as the need arises.

Close Relations of Program Evaluation

Many endeavors share similar qualities with program evaluation and may overlap with program evaluation. Individuals who develop skills as professional evaluators are well equipped to practice in these closely related areas and to leverage these activities to supplement a program evaluation.

Auditing is a formal process in which an independent party determines the extent to which an activity is performed in accordance with specified procedures (Schwandt, 2005). The U.S. Government Accountability Office (GAO) defines performance auditing in a way that is nearly indistinguishable from program evaluation. According to the GAO (2021), performance audits are "engagements that provide objective analysis, findings, and conclusions to assist management and those charged with governance and oversight to, among other things, improve program performance and operations, reduce costs, facilitate decision making by parties with responsibility to oversee or initiate corrective action, and contribute to public accountability" (p. 217).

Implementation science is an emerging field that studies how to increase the uptake of evidence-based practices by practitioners—mainly in health, human service, and education. The evidence-based practices of interest may have been determined through systematic program evaluation or research (Bauer, 2015; Eccles, 2006). Implementation science picks up where evaluation leaves off, with a focus on how to influence health, education, and human service practitioners to adopt proven practices for the benefit of the populations they serve.

Improvement science is a systematic approach for enhancing quality based on evidence. More specifically, it is "a data-driven change process that aims to systematically design, test, implement, and scale change toward systemic improvement, as informed and defined by the experience and knowledge of subject matter experts" (Lemire et al., 2017, p. 32). The use of evaluation results for program improvement is usually desired. But, in evaluation, program improvement is not the sole focus as it is in improvement science.

Monitoring involves continuous data collection on key indicators of program progress and implementation. In the context of international development, *monitoring and evaluation* or *M&E* refers to a range of activities that involve regular data collection about program activities and assessment of program quality. Data gathered for monitoring purposes may be used later for evaluation.

Performance measurement is the routine collection and reporting of data related to program implementation and results (i.e., performance), typically by or for government agencies. In the United States, the Government Performance and Results Act and the Program Assessment Rating Tool are examples of performance measurement initiatives (Davies et al., 2006).

Policy analysis is the study of the impact of public policies. Policies are intended to bring about particular outcomes or changes, but unlike programs, they typically do not involve specific services or activities. Thus, policy analysis is somewhat different from program evaluation and typically occurs on a much larger scale.

Quality assurance and *quality control* are processes for ensuring established quality standards are met, such as in manufacturing (Williams, 2005). Evaluative activities for ensuring the quality of processes and products abound in the business sector; they go by many names, such as *total quality management, continuous quality improvement, quality circles*, and *Six Sigma*.

Examples of Program Evaluation in Various Contexts

Below are some examples of purposes served by program evaluations in education, public health, social services, and business. These examples identify possible areas of focus and intended uses of evaluation in each field, illustrating the breadth of program evaluation activities.

Education

1. To determine the value of a middle school's block scheduling
2. To satisfy an external funding agency's demands for reports on the effectiveness of an afterschool program it supports
3. To determine the effectiveness of restorative justice practices as an alternative to disciplining students through suspension and expulsion

Public Health

4. To inform the development of a media campaign to promote breast cancer screening
5. To determine the effectiveness of an outreach program on infant immunization rates
6. To identify ways to improve a program designed to increase the diversity of the public health workforce

Social Services

7. To assess the costs and benefits of a job training program
8. To decide whether to modify a program to facilitate the transition of individuals who have experienced chronic homelessness into stable housing
9. To determine the impact of a prison's early-release program on recidivism

Business and Industry

10. To judge the effectiveness of a corporate training program for improving teamwork
11. To determine the effect of a new flextime policy on productivity, recruitment, and retention
12. To recommend ways to improve retention among employees who are people of color

As should be apparent at this point, the underlying principles, general focus, and potential uses of an evaluation are portable from one program evaluation context to another. Evaluation can inform decisions about a corporate media campaign, a community wellness program, or a school district's student assessment system. It can help build organizational capacity at a tech company, state department of education, or county health department. Evaluation can reveal ways to improve educational programs at rural and urban school districts and large and small colleges. It can inform decisions about programs operated by vocational education centers, community mental health clinics, university medical schools, or county cooperative extension offices. Such examples could be multiplied ad infinitum, but these should suffice to make our point.

Roles and Activities of Professional Evaluators

Professional evaluators play numerous roles and conduct multiple activities in carrying out their work:

- Evaluators support program planners to develop and design programs by (a) conducting needs assessments to determine what services or activities are most needed by a particular group and (b) helping to create logic models and theories of change to describe how a program will bring about change.
- Evaluators help organizations learn and grow by facilitating critical reflection, training staff to collect and use data to inform decision-making, and developing internal evaluation capacity.
- Evaluators serve as "critical friends" who point out weaknesses and vulnerabilities so organizations can fix them before they become serious problems.
- Evaluators provide technical expertise to organizations, making stakeholders aware of evidence that relates to their programmatic work.
- Evaluators may act as advocates and change agents, speaking truth to power to shed light on inequities and injustices and amplify the voices of those who have been marginalized.
- Evaluators are sometimes watchdogs who ensure that organizations use their resources responsibly and ethically.

Thus, evaluators take on many roles. In noting the tension between advocacy and neutrality, Weiss (1998b) wrote that the role(s) evaluators play depends

heavily on an evaluation's context. An evaluator may serve as a teacher or critical friend in an evaluation designed to improve a new reading program. An evaluator may act as a facilitator or collaborator with a community group appointed to explore ways to reduce food insecurity. In evaluating a program to enhance the employability of immigrants in a state, an evaluator may help stimulate dialogue among immigrants, policymakers, and nonimmigrant groups. Finally, an evaluator may serve as an outside expert in a study for Congress on the effectiveness of annual testing in improving student learning.

In carrying out these roles, evaluators undertake many activities, such as:

- negotiating with stakeholder groups to define the purpose of evaluation
- developing contracts, hiring and overseeing staff, and managing budgets
- identifying disenfranchised or underrepresented groups
- working with advisory panels
- collecting, analyzing, and interpreting qualitative and quantitative data
- communicating frequently with various stakeholders to seek input on the evaluation and to report results
- writing and disseminating reports
- meeting with the press and other representatives to report on progress and results
- recruiting other evaluation experts to evaluate their evaluations.

These and many other activities make up the work of professional evaluators. Professional evaluators are formally trained and educated in evaluation, attend professional conferences, read widely in the field, and identify professionally as evaluators.

Professional evaluators may be internal to the organizations whose programs they evaluate, or they may be external. Whether an evaluation is internal or external denotes how the evaluator is employed—as a regular staff member of the organization whose program is being evaluated (internal) or as a contracted consultant (external). Their employment status does not affect the kinds of evaluations conducted, their rigor or importance, or the evaluators' professional obligations.

Chapter 3 includes a detailed overview of the competencies that professional evaluators should possess for this varied and challenging work.

The Promise and Limitations of Program Evaluation

Program evaluation is a systematic process to collect, analyze, and interpret sound evidence to judge program quality and inform decision-making. When policymakers value science and evidence, they call for evaluations to inform them about what's working and what's not in the best interests of the people, organizations, and communities they serve. Within organizations, decision-makers and program staff can use evaluations to plan and improve programs. Members of the public

can use evaluation reports to learn about the impact of programs funded by their tax dollars and make choices about the programs, organizations, and institutions they will patronize. *This is the promise of program evaluation*.

In contrast, for policymakers and organizational decision-makers driven by ideology, politics, or personal preference, even the most rigorous evaluation will have little influence. Such actors may still request and pay for evaluations. But in these cases evaluations are merely symbolic—to give the impression that decisions, policies, and programs are based on science, evidence, and rationality. Or, worse, evaluators or evaluations may be manipulated to produce results that reinforce existing positions.

In truth, most evaluators work in the middle ground of these extremes. Some people embrace evaluation. Others may see it as a threat to what they care about or as a disruption to what is known and comfortable. It can be difficult for people to question the work they do or support. That is what evaluative inquiry prompts one to do. Evaluation has a vital role to play in a healthy democracy. Like democracy, it is imperfect—in both its execution and its effects. Many factors other than evidence from disciplined inquiry influence decision-making. In a democracy, elected and appointed officials must attend to many issues. Results of evaluations are not their sole source of information, nor should they be. They must consider citizens' input and expectations, as well as resource limitations. Evaluation is one source of information that influences policymakers, organizational leaders, and individual consumers.

The promise of evaluation is not limited to the information it generates about specific programs. Schwandt (2008) argued for an "intelligent belief in evaluation" (p. 139). He noted that evaluation is about not just its methods or findings, but a way of thinking that links reasoning with evidence. This intelligent belief in evaluation is critical in an "experimenting society":

> This is a society in which we ask serious and important questions about what kind of society we should have and what directions we should take. This is a social environment indelibly marked by uncertainty, ambiguity, and interpretability. Evaluation in such an environment is a kind of social conscience; it involves serious questioning of social direction; and it is a risky undertaking in which we endeavor to find out not simply whether what we are doing is a good thing but also what we do not know about what we are doing. (p. 143)

As you learn about the theory and practice of evaluation, we hope you will also develop your own "intelligent belief in evaluation." It will help you navigate an uncertain and ever-changing world.

Key Points

1. Evaluation—judging the quality or value of something—is an everyday human activity. *Program evaluation* is a systematic process to determine the merit, worth, or significance of a coordinated set of activities designed to achieve specific purposes.

2. The main purposes of evaluation are to provide information that program decision-makers and other stakeholders can use to (1) design or transform their programs (through developmental evaluation); (2) improve their programs while they are underway (through formative evaluation); and (3) determine the overall merit, worth, and significance of programs (through summative evaluation).
3. Most evaluations focus on a program's processes and outcomes. Some evaluations focus on program costs and the value of the program's outcomes relative to those costs.
4. Evaluation and research seem similar because they use some of the same methods for data collection and analysis. Evaluation has a practical orientation to providing useful information to stakeholders. Research is typically undertaken to advance knowledge within a discipline.
5. Other forms of inquiry—such as auditing, implementation science, improvement science, monitoring, performance measurement, policy analysis, and quality assurance—share similar characteristics with evaluation. Evaluators can learn from the theories and strategies used in these fields, and vice versa.
6. Evaluation can serve individuals, policymakers, organizations, and society at large by providing useful and credible information to inform decision-making about what works and where resources should be invested. However, evaluation is just one source of information.

Discussion Questions

1. How does the discussion of evaluation in this chapter compare with how you previously thought about evaluation?
2. Think of an example when you evaluated something very informally. How did that experience differ from formal evaluation as described in this chapter?

Application Exercises

1. Search the web for a credible news story that mentions a "program evaluation." (Be sure to search for *program evaluation* and not just *evaluation*, since that term on its own is used to mean many different things.) Answer as many of the following questions as possible:
   a. What was evaluated?
   b. What prompted the evaluation?
   c. Was there any controversy about the evaluation? If so, what were the contentious issues?
   d. Were any significant decisions impacted by the evaluation? Who was involved in those decisions? What groups were affected?
   e. Who conducted the evaluation? Is there any information about the qualifications or credibility of the evaluator(s)?

2. Think of a program you are familiar with that would benefit from a formal evaluation.
   a. Why did you pick this program?
   b. Who would use information from this evaluation? How would (or should) they use the information?
   c. What would be the primary purpose of this evaluation—developmental, formative, or summative?
   d. What aspect of the program would the evaluation focus on—its processes, outcomes, or costs?
   e. How might the evaluation affect the organization that operates the program? The people or conditions served by the program?

Suggested Resources

American Evaluation Association. (n.d.). *What is evaluation?* https://www.eval.org/About/What-is-Evaluation
This page on the American Evaluation Association's website features a formal statement on the definition of evaluation and videos of practicing evaluators talking about what evaluation is.

Greene, J. (2017, April 12). *What is evaluation?* [Video]. YouTube. https://youtu.be/CvmzJXI5xwQ
In this seven-minute video, noted evaluation scholar Jennifer Greene discusses several concepts that are central to the theory and practice of evaluation.

LaVelle, J. (2010). Describing evaluation. *AEA365: A Tip-a-Day by and for Evaluators*. https://aea365.org/blog/john-lavelle-on-describing-evaluation/
This blog post includes a useful graphic to help conceptualize the similarities and differences between research and evaluation.

Schwandt, T. (2008). Educating for intelligent belief in evaluation. *American Journal of Evaluation*, *29*(2), 139–150.
In this article, first presented as a plenary speech at the 2007 American Evaluation Association conference, noted evaluation scholar Tom Schwandt argued that an "intelligent belief in evaluation" is critical for advancing society.

2

Origins and Development of Program Evaluation as a Discipline and Profession

Orienting Questions

1. What were the characteristics of evaluation before the field began to take shape as a distinct profession and form of inquiry?
2. What were the major periods in the development of professional program evaluation? What events and developments are associated with each period?
3. What events spurred the emergence of modern program evaluation?
4. What other major developments helped advance and shape evaluation as a discipline and profession?

In this chapter, we review the history of evaluation and its progress toward becoming a full-fledged profession and distinct discipline. This history illuminates the forces that have helped shape the field of program evaluation and key advancements in its growth and maturation. This history is United States-centric for two reasons. First, the story of professional program evaluation begins in the United States. Second, as evaluation practitioners and teachers living and working in the U.S., this is the history we (the authors) know best. As we discuss later in the chapter, evaluation is spreading rapidly across the world. We leave the telling of the story of the evaluation's development outside of the United States to our international colleagues.

As Scriven (1996) noted, "Evaluation is a very young discipline—although it is a very old practice" (p. 395). He pointed out that the formal evaluation of industrial

crafts has taken place since prehistoric times: "As long as artifacts have existed—surely hundreds of millennia before any of the highly durable stone tools were made—it is very likely that craft workers have been evaluating their own and their fellow workers' products as part of a sustained development evaluation process" (Scriven, 2016, p. 35). Indeed, ever-advancing technological development by humans has been made possible by our ability to discern the strengths and weaknesses of products and processes and make improvements accordingly.

In the public sector, formal evaluation was evident as early as 2000 BCE, when Chinese officials conducted civil service examinations to measure the proficiency of applicants for government positions. Socrates used verbally mediated evaluations as part of the learning process. Centuries passed before formal evaluations gained a foothold in society to inform decision-making about how best to educate citizens and enhance their well-being.

The ascendancy of natural science in the 17th century was a necessary precursor to the premium that later came to be placed on direct observation of natural and social phenomena. Occasional tabulations of population size, mortality, and health grew into a fledgling form of empirical social research. In 1797, the word "statistics" appeared in *Encyclopedia Britannica*, which described it as "'a word lately introduced to express a view or survey of any kingdom, county, or parish'" (quoted in Louckx & Vanderstraeten, 2014, p. 530). Sociological historians Louckx and Vanderstraten (2014) noted, "These state-istics had to cover the growing need for information in the emerging 'enlightened' regimes and nation-states by means of surveys, population registers or censuses" (p. 530).

Quantitative surveys were not the only precursor to modern social research in the 1700s. Rossi and Freeman (2004) gave an example of a British sea captain who divided the crew into a treatment group that ate limes and a control group that did not. This experiment showed that eating limes prevented scurvy. As a result, British seafarers "were eventually compelled to consume citrus fruit regularly—a practice that gave rise to the still-popular term *limeys*" (p. 3).

These few examples demonstrate that as societies evolved, so did the need to systematically determine the status and cause of conditions and assess the impact of potential remedies to problems.

1800–1940: The Seeds of Modern Program Evaluation are Planted

In this section, we highlight key historical developments that set the stage for the development of program evaluation as a distinct form of inquiry.

Empirical Investigation of the Quality of Education Programs and Practices

The seeds for modern program evaluation were planted in the early 1800s with efforts to systematically assess school quality. In England, dissatisfaction with the education system spurred reform movements in which government-appointed

royal commissions heard testimony and used other less formal methods to evaluate the respective institutions. This led to still-existing systems of external inspectorates for schools in England and much of Europe (SICI, 2016; Standaert, 2004). In the United States, educational evaluation took a slightly different turn. It was influenced by Horace Mann's comprehensive annual, empirical reports on education in Massachusetts in the 1840s and the Boston School Committee's 1845 and 1846 use of printed tests in several subjects. This was the first instance of wide-scale student assessment and served as the basis for school comparisons. These two developments in Massachusetts were the first attempts to objectively measure student achievement to assess the quality of a large school system. They set a precedent for today's widespread use of student test scores as the primary means for judging school effectiveness.

In the late 1800s, liberal education reformer Joseph Rice conducted one of the first comparative studies to assess the quality of instructional methods. His goal was to substantiate his claims that school time was used inefficiently. To do so, he compared schools that varied in the amount of time spent on spelling drills and then examined the students' spelling ability. He found negligible differences in students' spelling performance between schools where students spent as much as 100 minutes per week on spelling instruction and those where they spent as little as 10 minutes per week. He used these data to encourage educators to scrutinize their practices empirically. Rice's study is considered the "first formal educational evaluation in the United States" (Stufflebeam & Coryn, 2014, p. 30).

A landmark evaluation from the early 1900s was Flexner's (1910) review of medical schools. Backed by the American Medical Association and the Carnegie Foundation, he assessed 155 medical schools operating in the United States and Canada. Following a series of one-day site visits to each school by himself and one colleague, Flexner delivered scathing reviews of the schools' quality and the state of medical education in general. He was not deterred by lawsuits or death threats due to what the medical schools viewed as his "pitiless exposure" (p. 87) of their medical training practices. He delivered his evaluation findings in scathing terms. He called Chicago's 15 medical schools "the plague spot of the country in respect to medical education" (p. 84). Soon "schools collapsed to the right and left, usually without a murmur" (p. 87). Flexner's reports were unambiguously evaluative. His review was a precursor to formal accreditation of academic programs in higher education.

The educational testing movement gained momentum in the early 1900s as measurement technology made rapid advances under E. L. Thorndike and his students. A pioneer in educational testing, Thorndike established norms for student performance in math, reading, handwriting, and other subjects. These norms enabled school administrators to compare their students' knowledge and abilities with the average achievement of a representative sample of children. By 1918, objective testing was flourishing, pervading the military, private industry, and all levels of education. The 1920s saw the rapid emergence of norm-referenced tests, which were designed to rank students. By the mid-1930s, more than half of U.S. states had some form of statewide testing. During this period, educators regarded measurement and evaluation as nearly synonymous. The latter was usually thought of as summarizing student test performance and assigning grades.

Although program evaluation as we know it today was still in its infancy, useful measurement tools for evaluation were proliferating. Very few meaningful, formal evaluations of school programs or curricula were published during this period. One notable exception was the ambitious, landmark Eight-Year Study (Smith & Tyler, 1942). The Eight-Year Study set a new standard for educational evaluation with its sophisticated methodology for measuring learning outcomes. With this and later studies Tyler (e.g., 1950) also planted the seeds of standards-based testing—that is, testing to determine students' mastery of subject matter they were expected to learn, without regard for ranking or comparison with peers as in norm-referenced testing. (In Chapter 6, we discuss Tyler's profound impact on program evaluation, especially in education.)

Beginning in the late 1930s (when positions of influence in federal agencies and higher education were largely held by White men), Black education researchers began to shed light on racial inequities in education (Thomas & Campbell, 2021). Ambrose Caliver, the first Black person to receive a Ph.D. from Columbia University, had an influential role in the U.S. Office of Education beginning in 1932. He spearheaded the collection, analysis, and dissemination of data that revealed gross inequities in the U.S. education system based on race (Thomas & Campbell, 2021; Hood, 2001). Reid Jackson investigated the quality of schools for Black children within the segregated education systems in Kentucky, Florida, and Alabama. Jackson, who held multiple administrator and faculty positions at historically Black colleges and universities, broke new ground by approaching his studies through the lens of culture and race (Thomas & Campbell, 2021; Hopson & Hood, 2005). The launch of *The Journal of Negro Education* in 1932 provided an important outlet for the publication of evaluative studies that focused on the role of race in education (Hood, 2001; Thomas & Campbell, 2021).

Growth of Applied Social Research

In the early 1900s, foundations for evaluation were also being laid in fields beyond education, including health, human services, and business. For example, Cronbach and his colleagues (1980) cited surveys of slum conditions, management and efficiency studies in schools, and investigations of local government corruption. Rossi, Freeman, and Lipsey (2004) noted that at that time evaluations were being conducted in the field of public health, where studies focused on assessing efforts to control infectious diseases. Fredrick Taylor's influential scientific management theory focused on discovering the most efficient way to perform a task and then training all staff to perform it that way. The emergence of "efficiency experts" in industry soon permeated the business community. As Cronbach et al. (1980) noted, "business executives sitting on the governing boards of social service agencies pressed for greater efficiency in those services" (p. 27). Some cities and social service agencies began to develop internal research units. Social scientists began to trickle into government service. They conducted applied social research in areas such as public health, housing, and work productivity. However, these social research precursors to evaluation were small, isolated activities. They had little impact on the lives of the citizenry or the decisions of government agencies.

With the Great Depression in the 1930s came the sudden proliferation of government services and agencies as President Roosevelt's New Deal programs were implemented to salvage the U.S. economy. This was the first major growth in the U.S. federal government in the 20th century, and its impact was profound. Federal agencies were established to oversee new national programs in welfare, public works, labor management, urban development, health, education, and numerous other human service areas. Increasing numbers of social scientists went to work in these agencies. Applied social research opportunities abounded. Social science academics began to join with their agency-based colleagues to study a variety of variables related to these programs. While some scientists called for explicit evaluation of the new social programs (e.g., Stephan, 1935), most pursued applied research at the intersection of an agency's needs and their personal interests. Thus, academic sociologists pursued questions that were of interest to both the discipline of sociology and the agency. However, these questions often originated with the academics. The same trend occurred with economists, political scientists, and other academics who conducted research on federal programs. Their projects were considered to be "field research" and provided opportunities to address important questions within their disciplines.

1941–1963: Applied Social and Educational Research Become Commonplace

This period was not especially remarkable in terms of the development of program evaluation. It is notable, however, that applied social research and education research became more commonplace and, in a few instances, began to be institutionalized. However, the limitations of using applied research to answer pressing questions about program quality, value, and effectiveness became apparent. This situation set the stage for the new discipline of evaluation to emerge.

Applied social research expanded during World War II as researchers investigated programs to help military personnel determine how to reduce vulnerability to propaganda, increase morale, and improve the training and job placement of soldiers. In the following decade, studies focused on new government programs for job training, housing, family planning, and community development. As in the past, such studies tended to focus on particular facets of the program in which the researchers happened to be most interested. As these programs expanded, however, social scientists began to broaden their studies to examine entire programs.

With this broader focus came more frequent use of social research methods to investigate and improve specific programs. Rossi et al. (2004) stated that it was typical during this period for social scientists to be "engaged in assessments of delinquency-prevention programs, psychotherapeutic and psychopharmacological treatments, public housing programs, educational activities, community organization initiatives, and numerous other initiatives" (p. 23). Such work also spread to other countries and continents. Many countries in Central America and Africa were the sites of evaluations examining health and nutrition, family planning,

and rural community development. Most of these studies relied on existing social research methods and did not extend the conceptual or methodological boundaries of evaluation beyond those already established for behavioral and social research. Such efforts would come later.

Developments in the education sector in the 1940s through early 1960s were unfolding somewhat differently. In this period, earlier developments in educational evaluation were consolidated and refined. School personnel worked to improve standardized testing, quasi-experimental design research, accreditation, and school surveys. In the 1950s and early 1960s, there were also efforts to enhance the Tylerian approach to evaluation (see Chapter 6) by teaching educators how to state objectives in explicit, measurable terms so that student progress toward these objectives could be validly and reliably measured.

Black evaluators and researchers continued to call attention to racial inequities in education. Ambrose Caliver's influence in the U.S. Office of Education continued throughout the 1950s (Thomas & Campbell, 2021; Hood, 2001). Aaron Brown evaluated the quality of accredited secondary schools for Black children in the South. Brown, a student of Tyler's, used criteria established by six regional educational associations to evaluate 93 schools (Brown, 1944). He identified both "satisfactory features" and "weaknesses" across the schools and discussed possible reasons for notably high and low scores on various criteria (p. 493). Many of the possible causal or interacting factors he described were "peculiar to schools operating for Negroes" (p. 494). Thus, while it would be many decades before culturally responsive practices gained prominence in the evaluation field, Brown's work called attention to the need to consider culture and context when interpreting the results of educational research and evaluation. Leander Boykin, the first Black person to attain a Ph.D. in education at Stanford University (Hood, 2001), studied the differences in financial resources and teacher salaries in schools for Black and White children in the south. Boykin argued for using mixed methods in evaluation, using evaluation to improve education programs, involving stakeholders in the evaluation process, and considering the social and economic context of education programs (Thomas & Campbell, 2021; Hood, 2001). Like Brown's, Boykin's work foreshadowed later developments in the field of program evaluation.

In 1957, the Soviets' successful launch of Sputnik I sent tremors through the U.S. establishment that were quickly amplified into calls for more effective teaching of math and science to students in U.S. schools. The reaction was immediate. The National Science Foundation started to fund science and math curriculum development initiatives, along with the evaluation of those efforts. According to Cronbach et al. (1980), these studies "were sometimes simple and rather informative, but a few were extensive and conformed to the canons of experimental design" (p. 31). Passage of the National Defense Education Act (NDEA) of 1958 poured millions of dollars into massive, new curriculum development projects, especially in math and science. Only a few projects were funded, but their size and perceived importance led policymakers to fund evaluations of most of them.

Notwithstanding the expanding program evaluation efforts, theoretical work related directly to evaluation did not exist. Therefore, those who conducted

evaluation studies were left to utilize what they could from applied social, behavioral, and educational research. Their gleanings were so meager that Cronbach (1963) penned a seminal article in which he sharply criticized past educational evaluations. He called for evaluators to move beyond "comparing score averages" to assessing far-ranging outcomes, including "attitudes, career choices, general understandings and intellectual powers, and aptitude for further learning in the field" (p. 247). His recommendations had little immediate impact. However, they did catch the other education scholars' attention, helping to spark a greatly expanded conception of evaluation that would emerge in the next decade.

1964–1969: Modern Program Evaluation Emerges

The developments discussed so far were not sufficient in themselves to launch a strong and enduring program evaluation movement. However, they did make conditions ripe for such a development. Much happened to spur the modernization of evaluation between 1964 and 1969—a brief but significant phase in the field's formation. Suddenly, the need for specialized approaches and professionals to conduct evaluations became acute. Critical developments in this period included (1) massive increases in U.S. federal spending on social programs and (2) a Congressional mandate to evaluate programs funded by the expansive Elementary and Secondary Education Act. Yet, there was a lack of scholarship and specialists in evaluation to address the growing demand.

Rapid Expansion of Federally Funded Social Programs

Conditions were ideal for accelerated conceptual and methodological development in evaluation, and a catalyst was found in initiatives spearheaded by U.S. President Lyndon Johnson, who took office in 1964. His "War on Poverty" legislation sought to equalize and enhance opportunities for all citizens in virtually every sector of society. He aimed to realize his vision for a "Great Society" by pouring millions of dollars into programs in education, health, housing, criminal justice, unemployment, urban renewal, and many other areas. Federal government spending on anti-poverty and other social programs increased by 600 percent after inflation from 1950 to 1979 (Bell, 1983). There was strong interest in learning how programs in areas such as job training, urban development, and housing were working. Managers and policymakers wanted to know how to improve their programs and which strategies worked best to achieve their ambitious goals. Congress wanted information on the types of programs that were worthy of continued funding. Increasingly, evaluations were mandated. In 1969, federal spending on grants and contracts for evaluation was $17 million. By 1972, it had expanded to $100 million (Shadish et al., 1991).

Unlike the private sector, where accountants, management consultants, and research and development experts were readily available to provide feedback

on corporate programs' productivity and profitability, these huge, new social investments had no similar mechanism in place to examine their progress. Some government employees had relevant competence—social scientists and technical specialists in the various federal departments, particularly in the General Accounting Office (GAO)—but they were too few and not sufficiently organized to determine the effectiveness of these vast government innovations. To complicate matters, many inquiry methodologies and management techniques that worked on smaller programs proved inadequate for programs of the size and scope of these sweeping social reforms.

For a time, it appeared that another concept developed and practiced successfully in business and industry might be successfully adapted for evaluating these federal programs: the Planning-Programming-Budgeting System (PPBS). PPBS was used by Ford Motor Company and later brought to the U.S. Department of Defense by Robert McNamara when he became President Kennedy's secretary of defense in 1961. The PPBS was a variant of the approaches used by many large aerospace, communications, and automotive companies. It was aimed at improving system efficiency, effectiveness, and budget allocation decisions by defining organizational objectives and linking them to system outputs and budgets. Many thought the PPBS would be ideally suited for the federal agencies charged with administering War on Poverty programs, but few bureaucrats heading those agencies were eager to embrace it. The stage was set for the creation of new evaluation approaches and methods, as well as a new kind of professional, with a somewhat different type of training and orientation, to apply them.

Elementary and Secondary Education Act (ESEA) of 1965

The single event that arguably was most responsible for the emergence of modern program evaluation is the passage by the U.S. Congress of the Elementary and Secondary Education Act (ESEA) of 1965. This event sent a shock wave through the U.S. education system, awakening both policymakers and practitioners to the importance of systematic evaluation. This bill proposed a considerable increase in federal funding for education, with tens of thousands of federal grants to local schools, state and regional agencies, and universities. The bill's largest component was Title I, destined to be the costliest federal education program in U.S. history. Wholey and White (1973) argued that Title I was the most important among the array of legislation that influenced evaluation at the time.

When Congress began its deliberations on the proposed ESEA, legislators, especially in the Senate, expressed concerns about a lack of convincing evidence that any federal funding for education had resulted in real educational improvements. Indeed, some members of Congress believed federal funds allocated to education prior to ESEA had sunk like stones into the morass of educational programs with scarcely a ripple to mark their passage. Robert F. Kennedy was the most persuasive voice among these. He insisted that the ESEA require each grant recipient to file an evaluation report showing what had resulted from the expenditure of

the federal funds. This Congressional evaluation mandate was ultimately approved for Title I (compensatory education) and Title III (innovative educational projects). The requirements "reflected the state-of-the-art in program evaluation at that time" (Stufflebeam et al., 2000, p. 13). These requirements reflected an astonishing amount of micromanagement at the Congressional level. They also heightened attention to accountability, calling for standardized testing to demonstrate student learning.

Some important milestone evaluation studies occurred at this time. These included the evaluations of Title I, Head Start (Westinghouse, 1969), and the *Sesame Street* television series (Ball & Bogatz, 1970). The evaluations of *Sesame Street* demonstrated some of the first uses of formative evaluation, as portions of the program were examined to provide feedback to program developers for improvement.

The passage of the ESEA in 1965 deserves its historical recognition as the birth of contemporary program evaluation. However, it was also marked by significant travail. Social and educational researchers lacked tools and frameworks to evaluate programs effectively. As the evaluation field was still in its infancy, few training programs were in place, and methodologies were largely borrowed from field studies in the social and behavioral sciences.

Emergence of Evaluation Specialists and Evaluation Approaches

The need for experts who could conduct useful and rigorous evaluations was sudden and pressing, and the market responded. Congress provided funding for universities to launch new graduate training programs in educational research and evaluation, including fellowship stipends for graduate study in those specializations. Several universities began graduate programs for educating evaluators. Political science programs spawned schools of public administration to train administrators to manage and oversee government programs. Policy analysis emerged as a means to assess different policy options and measure the impacts of implemented policies. Graduate education in the social sciences ballooned. The number of people completing doctoral degrees in economics, education, political science, psychology, and sociology grew from 2,845 to 9,463, an increase of 333%, from 1960 to 1970 (Shadish et al., 1991). Many graduates of these programs pursued careers evaluating programs in the public and nonprofit sectors.

Until the late 1960s, there was minimal theoretical and methodological guidance specific to evaluation to inform the work of these new evaluation practitioners. They were left to draw from theories in their primary disciplines and glean what they could from existing social research methods. Such methods included experimental design, psychometrics, survey research, and ethnography. In response to the need for more scholarship on evaluation, some important books and articles were published. Suchman (1967) wrote a book reviewing different evaluation methods, and Campbell (1969) argued for more social experimentation to examine program effectiveness. Campbell and Stanley's book (1966) on experimental and quasi-experimental designs influenced many working in evaluation to

adopt this approach. Scriven (1967), Stake (1967), and Stufflebeam (1968) began to write articles about evaluation practice and theories. Evaluation as a distinct form of inquiry began to take shape.

These conditions led to much excitement among evaluation pioneers about this new area of inquiry and their role in improving social conditions. Donald Campbell, the renowned research methodologist who trained several individuals who later became leaders in evaluation, wrote of the "experimenting society" in his article "Reforms as Experiments." He urged managers to use data collection and experiments to learn how to develop good programs (Campbell, 1969). He argued that managers should advocate not for their programs but for solutions to the problems their programs were designed to address. He suggested that by advocating for the development and testing of solutions, managers could help policymakers, citizens, and other stakeholders become more patient with the difficult process of solving social problems such as crime, unemployment, and illiteracy. As aptly put by Shadish, "There was this incredible enthusiasm and energy for social problem solving" during this period (Oral History Project Team, 2003, p. 271).

1970–1999: Program Evaluation Becomes a Profession

In the last 30 years of the 20th century, program evaluation matured substantially as a professional practice and distinct form of inquiry. There was substantial growth in the volume of evaluation-specific literature. Professional evaluation associations formed. Standards of practice were established to help shape the professional identity of evaluators. During this time, the contexts and approaches to evaluation diversified, with evaluation increasingly leveraged to enhance organizational learning.

Growth of Evaluation Literature and Professional Evaluation Associations

In the absence of any comprehensive textbooks on evaluation, Caro (1971) published a collection of readings on evaluation. Soon after, program evaluation textbooks began to be published. Examples include *Evaluation Research: Methods of Assessing Program Effectiveness* by Weiss (1972), *Educational Evaluation: Theory and Practice* by Worthen and Sanders (1973), and *Evaluation: A Systematic Approach* by Rossi, Freeman, and Rosenbaum (1979). Many more followed, including subsequent editions of those early texts. Articles about evaluation began to appear with increasing frequency in academic journals. These publications featured new evaluation models and approaches to respond to the needs of specific types of evaluation (e.g., ESEA Title III evaluations, and evaluations of mental health programs).

The number of journals that focused on evaluation grew dramatically in this period. These included *American Journal of Evaluation; Canadian Journal of Program Evaluation; Educational Evaluation and Policy Analysis; Evaluation: The International Journal of Theory, Research, and Practice; Evaluation and Program Planning; Evaluation and the Health Professions; Evaluation Practice; Evaluation Studies Review Annual;*

Evaluation Quarterly; Evaluation Review; ITEA Journal of Tests and Evaluation; New Directions for Program Evaluation; Performance Improvement Quarterly; Practical Assessment, Research, and Evaluation; Research Evaluation; and *Studies in Educational Evaluation*. Some journals omitted an explicit reference to evaluation in their titles but highlighted it in their contents. These included *Journal of Policy Analysis and Management*, *Performance Improvement Quarterly*, and *Policy Studies Review*. In addition, the *Journal of Negro Education* continued to publish evaluative studies related to the education of Black children and racial disparities in the education system. Most originated in North America, with a few in Europe.

In the late 1970s and throughout the 1980s, the publication of evaluation books, including textbooks, reference books, and even compendia and encyclopedias of evaluation, increased markedly. In response to the demands for guidance and the collective experience gained from practicing evaluation in the field, a body of specialized evaluation literature developed and expanded.

Simultaneously, professional associations and related organizations were formed. The American Educational Research Association's Division H was an initial focal point for professional activity in evaluation. Two national professional associations were founded in the United States that focused exclusively on evaluation: the Evaluation Network in 1975 and the Evaluation Research Society (ERS) in 1976. In 1986, these organizations merged to form the American Evaluation Association (AEA). Many local and regional AEA affiliates were created to offer options for more localized professional exchange related to evaluation.

With a growing literature, professional associations, and conferences where evaluators could exchange ideas with colleagues engaged in similar work, evaluation began to take shape as a distinct discipline and profession.

Development of Standards for Evaluation

In 1975, 12 professional associations concerned with evaluation in education came together to form the Joint Committee on Standards for Educational Evaluation. The Committee's charge was to develop standards that evaluators, evaluation clients, and evaluation consumers could use to judge the quality of program evaluations in education settings. In 1981, the Joint Committee published *Standards for Evaluations of Educational Programs, Projects, and Materials*. The initial 30 standards for evaluation were organized under the headings of utility, feasibility, propriety, and accuracy. That is, they called for evaluations to be useful, practical, ethical, and valid. The primacy of the standards for the utility of evaluations is notable. It signaled the profession's commitment to providing useful and relevant service and information to evaluation stakeholders. A second edition of the standards was published in 1994 and a third in 2011; the latest edition introduced a new domain of standards focused on evaluation accountability.

In 1982, the Evaluation Research Society (ERS) published its own standards for program evaluation. They distinguished their standards from the Joint Committee's by noting they were for program evaluation in any context, not just for educational programs (Evaluation Research Society Standards Committee, 1982). The 55 ERS standards were organized around five domains of evaluation

activity, including (1) formulation and negotiation, (2) structure and design, (3) data collection and interpretation, (4) communication and disclosure, and (5) utilization. When ERS merged with the Evaluation Network (ENet) in 1986 to form the American Evaluation Association, the new organization determined that it would formulate new guidance to replace the ERS standards.

These activities to develop and communicate a shared understanding of what constituted quality in program evaluation were critical in shaping evaluation as a profession and distinguishing it from applied social and educational research. (In Chapter 3, we discuss the Joint Committee standards, the AEA guiding principles, and other sets of standards, principles, criteria, and competencies for program evaluation in more detail.)

Shifts in the Market for Evaluation and Role of Evaluators

While the infrastructure for professional evaluation was being formed, the markets for evaluation were changing dramatically. Ronald Reagan's election as the U.S. president in 1980 led to a sharp decline in the number of federally funded evaluations. Instead, the federal government awarded block grants to states. States made their own decisions about spending and their own choices about evaluation requirements. However, the decline in evaluation at the federal level resulted in a healthy diversification of evaluation settings and approaches (Shadish et al., 1991). Reflecting on the contraction of funding at the national level, Worthen (1995) observed that evaluation became more commonplace among state agencies and local organizations, with a more authentic commitment to using the results. He noted, "Evaluation plays an increasingly important informational role as the level of the evaluation becomes more local, while evaluations at national levels typically continue to serve symbolic, non-informational functions" (p. 29).

Many state and local agencies began conducting their own evaluations, with less reliance on external experts. Foundations and other nonprofit organizations increased their attention to evaluation. Senge's (1990) book, *The Fifth Discipline*, spurred thinking about how organizations learn and change. This topic was highly relevant to evaluators since they were concerned with getting stakeholders in organizations to use evaluation information. In the education sector, evaluation continued to focus on student outcomes, measured by standardized testing. In other fields such as public administration, adult learning, and organizational management and change, there was a growing interest in leveraging evaluation to enhance organizational learning. *Evaluative Inquiry for Learning in Organizations* (Preskill & Torres, 1998) brought these concepts to the attention of evaluators. Evaluators began thinking more broadly about the role of evaluation in organizations and tasks evaluators should perform. Reichardt (1994) published an article reflecting on what the field had learned from evaluation practice, suggesting that evaluators should become more involved in the planning stages of programs. He argued that evaluators' skills might be more useful in a program's beginning stages rather than after it ended. Evaluators increasingly used logic models (see Chapter 6)

to determine an evaluation's focus and put that focus in an appropriate context. Engaging in logic model development helped program stakeholders to think about their programs evaluatively (Rogers & Williams, 2006).

In 1996, the United Way of America introduced a strategy for outcome measurement for the nonprofit agencies it funded. Their approach represented a convergence of the traditional focus on outcome evaluation with organizational learning. *Measuring Program Outcomes: A Practical Approach* (United Way of America, 1996) offered nonprofits a framework that included developing logic models to link inputs, activities, outputs, and outcomes; employing quantitative and repeated measures of outcomes; and using results for program improvement. The activities were not labeled as "evaluation" by United Way. Instead, they were characterized as "a modest effort simply to track outcomes" (Hendricks et al., 2008, p. 16). United Way's outcome focus was different from the outcome focus in the education sector in that (1) accountability was considered secondary to the purpose of program improvement and (2) expectations for measuring outcomes were generally more realistic than requirements for public-sector agencies. The United Way recognized that many nonprofit human service organizations lacked resources to conduct sophisticated evaluations of all outcomes. It also understood that engaging in evaluation could catalyze organizational learning and improvement. Nonprofit organizations, like many public-sector organizations, had typically reported inputs and activities to funders, with minimal attention to outcomes. The move to assess and monitor program outcomes was a notable shift to more comprehensive evaluation of nonprofit programs. Furthermore, it demonstrated that evaluation of outcomes and evaluation for organizational learning were indeed compatible.

Diversification of Evaluation Approaches and Methods

During this period, several prominent writers in the field proposed new and divergent approaches to evaluation. Evaluation moved beyond simply measuring whether objectives were attained. Evaluators started to consider program managers' information needs. They realized that evaluation should address unintended outcomes as well as those that were intended. The importance of making judgments about merit and worth, not just goal achievement, became apparent. The role of values and standards in reaching those judgments gained attention among evaluation scholars and practitioners.

These new and controversial ideas spawned dialogue and debate that informed a developing evaluation lexicon and literature. Scriven (1972) worked to push evaluators beyond the rote application of objectives-based evaluation. To this end, he proposed goal-free evaluation, urging evaluators to identify and assess all program outcomes, whether intended or not. Stufflebeam (1971), responding to the need for evaluations that were more useful to decision-makers, developed the context, input, process, and product (CIPP) model. Stake (1975) proposed a responsive evaluation style that featured a high degree of interaction between

evaluators and stakeholders. Guba and Lincoln (1981) built on Stake's qualitative work. They proposed naturalistic evaluation, which led to much debate over the relative merits of qualitative and quantitative methods.

These approaches were dramatically different from the dominant experimental, social science paradigms. Collectively, these new ideas about evaluation greatly broadened earlier views. It became apparent that good program evaluation encompassed much more than simple application of the research skills. (We provide more details on these approaches in Part 2 of this book.)

As evaluation funders and practitioners diversified, the nature and methods of evaluation adapted and changed. Formative evaluations provided feedback for incremental change and improvement. Evaluators helped programs theorize and measure links between program actions and outcomes. Patton developed his utilization-focused evaluation approach, which emphasized the importance of identifying intended evaluation users and adapting questions and methods to those users' needs (Patton, 1975, 1978, 1986). Guba and Lincoln (1981) urged evaluators to make greater use of qualitative methods to develop "thick descriptions" of programs, providing more authentic portrayals of the nature of programs in action. Fetterman (1984) advocated for the use of ethnographic methods for educational evaluation. As different types of organizations funded more evaluations and expressed different needs, evaluators who had previously focused on policymakers (e.g., Congress, cabinet-level departments, legislators) as their primary audience began to consider multiple stakeholders and use more qualitative methods. Participatory methods for involving many different stakeholders, including those often detached from decision making, became commonplace. Although the dramatic decline in federal funding for evaluation caused anxiety among evaluators at the time, it actually increased the number, depth, and breadth of approaches to evaluation.

This burgeoning body of evaluation literature revealed sharp differences in the authors' philosophical and methodological preferences. It also underscored a fact about which there was much agreement: Evaluation was a multidimensional technical and political enterprise that called for new theories and methods. Shadish and his colleagues (1991) said it well when they noted, "As evaluation matured, its theory took on its own special character that resulted from the interplay among problems uncovered by practitioners, the solutions they tried, and traditions of the academic discipline of each evaluator" (p. 31).

21st Century Program Evaluation: 2000–Present

Today, evaluations are conducted in diverse settings using a variety of approaches and methods. The continued development of the field in the 21st century has involved rapid growth of evaluation around the globe, diffusion and mainstreaming of evaluation, proliferation of opportunities to learn about evaluation, development of evaluation competency frameworks, creation of systems for credentialing evaluators in Canada and Japan, issuance of policy statements by the American

Evaluation Association, and increasing attention to culture, race, and social justice in evaluation.

Rapid Global Expansion

Evaluation has grown rapidly worldwide since the turn of the century, spurred by international collaborative efforts to enhance the practice and highlight the importance of evaluation for improving the human condition.

Rist's (1990) study of differences in evaluation across countries identified the United States, Canada, Germany, and Sweden as countries in the "first wave of evaluation development." These were countries where modern program evaluation originated in the 1960s and 1970s. In this first wave, evaluation was oriented to improving social programs and interventions. Countries included in the "second wave" are the United Kingdom, the Netherlands, Denmark, and France. In these countries, evaluation began as an effort to control federal budgets and reduce government spending. Evaluation was oriented more to accountability and identifying unproductive programs than to social experimentation and program improvement.

A follow-up study to Rist's has not yet been conducted to identify specific countries in the third or subsequent waves of evaluation development, but there has been rapid growth of evaluation in Africa, Asia, and South America. Based on casual observation rather than systematic study, it appears that the expansion of evaluation in the global south was spurred in part by the demand for evaluation of international development efforts. Donors and multilateral agencies that were funding development projects wanted evidence of the value and impact of their investments. This led to a growing need for professionals in those locations who could conduct evaluations. Rather than relying solely on evaluators from North America and Europe, organizations and countries in the global south recognized and responded to a need to develop local expertise in evaluation.

In 2003, the International Organization for Cooperation in Evaluation (IOCE) was created. It began as a coalition of 24 voluntary organizations for professional evaluation (VOPEs) to "build and strengthen the global network of relationships between existing and emerging VOPEs." (IOCE, n.d., n.p.). By 2016, the official count of VOPEs grew to 173, representing about 52,000 individual members worldwide (IOCE, 2016).

Less than a decade after its formation, IOCE, along with UNICEF, helped facilitate the creation of EvalPartners in 2012. EvalPartners describes itself as an "innovative partnership whose members are Civil Society Organizations (CSOs) and Voluntary Organizations for Professional Evaluation (VOPEs)" (EvalPartners, 2021, para. 1). It works to enhance the capacity of organizations within civil society to engage in "national evaluation processes, contributing to improved country-led evaluation systems and policies that are equity-focused and gender equality responsive" (para. 8). In addition to creating the first-ever global forum for knowledge sharing about evaluation, a major achievement of EvalPartners was its work to designate 2015 as the International Year of Evaluation. EvalYear 2015,

as it came to be known, was endorsed by the United Nations General Assembly in Resolution 69/237: Building Capacity for the Evaluation of Development Activities at the Country Level (UN, 2015a). A UN press release described this resolution as follows:

> For the first time in the history of the United Nations, a landmark, stand-alone UN Resolution on national evaluation capacity development has been adopted by the Second Committee of the United Nations General Assembly. This resolution was presented to the Committee by Fiji, and supported at global and country level by the United Nations Evaluation Group (UNEG) and national evaluation partners around the world. It received a very strong cross-regional sponsorship from more than 42 countries and a general consensus recognizing evaluation capacity as a country-level tool to strengthen evidence-based policymaking (UN Web TV, 2014).

The UN sees evaluation as playing a critical role in achieving its sustainable development goals, which span 17 areas that need to be addressed in order to "end poverty, protect the planet, and ensure prosperity for all" (UN, n.d.).

Along with practice, evaluation scholarship has expanded far beyond evaluation's origins in Canada and the U.S. Examples of evaluation-focused journals that started in the 2000s include the *Evaluation Journal of Australasia*, which started in 2001; *Evidence Base*, launched by the Australia and New Zealand School of Government in 2012; and the *African Evaluation Journal* and the *International Journal of Evaluation and Research in Education*, both of which started in 2013.

Diffusion of Evaluation Practice

As evaluation has spread geographically, the practice has also become more widespread in terms of the individuals engaged in evaluation and evaluation-related tasks. In the nonprofit sector, it is common for in-house evaluators to have responsibility for major components of data collection and evaluation in their organizations. Individuals charged with evaluation are typically program managers and staff who have other program responsibilities. In 2003, Christie found that many of the evaluators she surveyed in California were internal and held other responsibilities, which were mostly management-related. Many had little or no training in evaluation and were unfamiliar with evaluation theories and approaches. Although we don't know of other, more recent studies that reveal how pervasive this situation is, the plethora of basic evaluation guides and toolkits online suggests that many foundations and nonprofit organizations expect their grantees and employees to engage in evaluation with little or no formal training.

For some, the diffusion of evaluation responsibilities within organizations may raise concerns about the quality of evaluation studies. We believe the involvement of more individuals in evaluation endeavors has great advantages. In his American Evaluation Association presidential address in 2002, James Sanders (a coauthor of this book) advocated for mainstreaming evaluation, which he described as "the process of making evaluation an integral part of an organization's