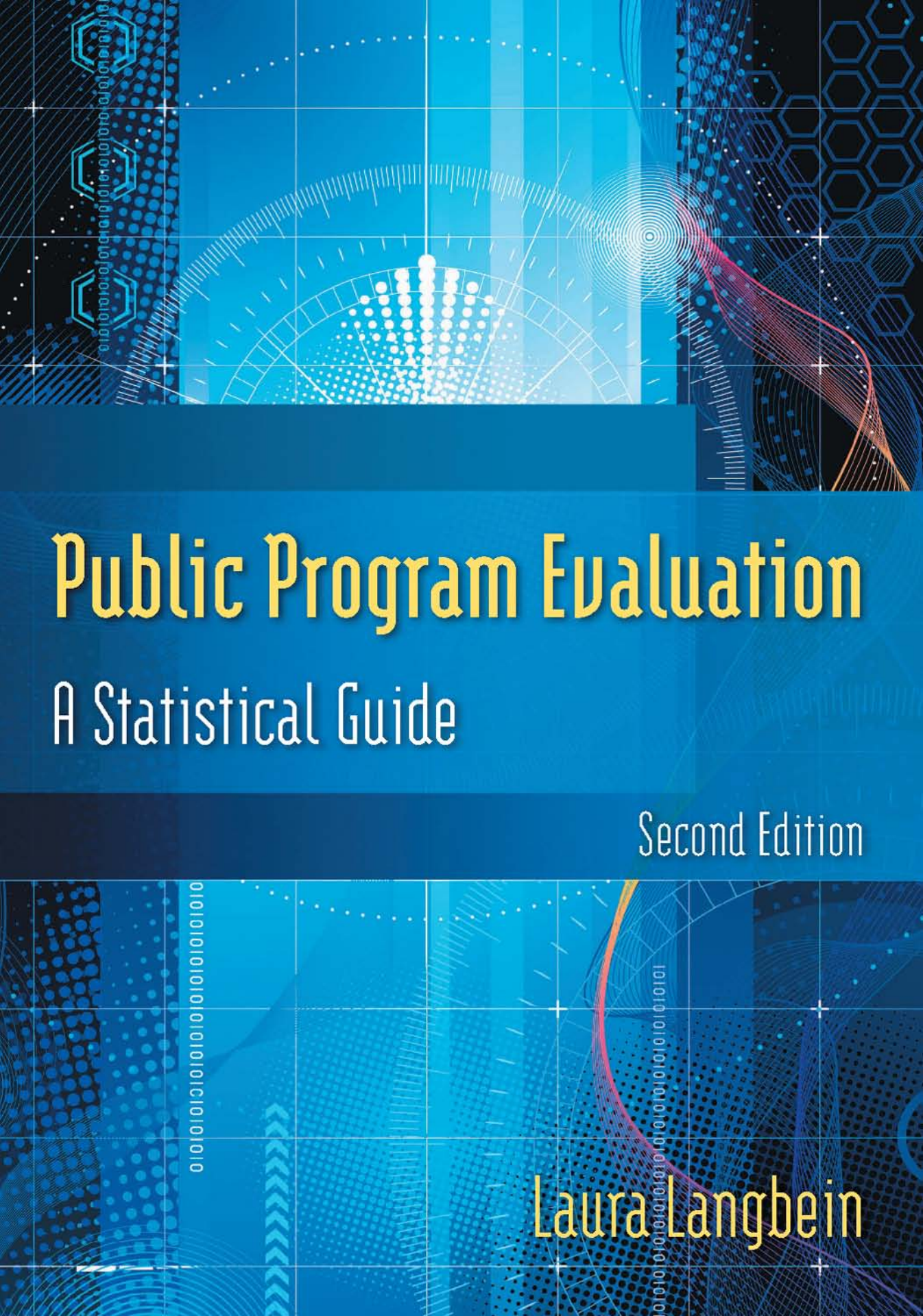

Public Program Evaluation
A Statistical Guide
Second Edition
Laura Langbein

Public Program Evaluation

This page intentionally left blank

Public Program Evaluation

A Statistical Guide

Second Edition

Laura Langbein

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2012 by M.E. Sharpe

Published 2015 by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN
711 Third Avenue, New York, NY 10017, USA

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

Library of Congress Cataloging-in-Publication Data

Langbein, Laura Irwin.
Public program evaluation : a statistical guide / by Laura Langbein. — 2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-7656-2612-7 (pbk. : alk. paper)
1. Policy sciences—Statistical methods. 2. Evaluation research (Social action programs)—Statistical methods. I. Title.

H97.L35 2012
352.3′5—dc23 2011045685

ISBN 13: 9780765626127 (pbk)

Contents

This page intentionally left blank

Preface

I have been teaching a course on program evaluation to public affairs graduate students for many years. That course introduces students to the conscious choice of defensible and feasible research designs so that they can evaluate whether public programs are doing what they are supposed to do and not doing what they are not supposed to do. Students enter the course after having taken one or more statistics courses. I ask them to select an appropriate, feasible research design and corresponding statistical tools in order to examine whether a particular program of interest is meeting its intended goals (and avoiding unintended consequences). The course focuses on the hardest (and, arguably, the most important) question of program evaluation: Did the outcome result from the program or would it have occurred anyway?

While questions of causal inference are among the most abstract in science, they nevertheless have to be answered in the real world, because program managers, legislators, and funding organizations legitimately want answers to these questions. Thus, the methods of program evaluation must be practical and as valid (or defensible) as possible. The consequence is that program evaluators cannot always use the scientifically ideal design of the randomized experiment to study program effects. Perhaps evaluators should not use randomized experiments very much because in some respects the design may not be ideal—and even ideal designs go awry. What then? There are many other research designs and a great deal of data. Accordingly, there are many research design textbooks—not only in the specialty of program evaluation but also in many other disciplinary fields, especially political science and sociology—and even more statistics (and econometrics) textbooks. But how do you put the two together? Once you have decided that an interrupted time-series design (or some other design) is a practical and reasonably defensible design to analyze the effects of a particular program, no text tells you which statistical techniques are appropriate for that design. The research designs are usually portrayed as pictures, graphs, or tic-tac-toe diagrams with Xs and Os. There is no statistical guide in the research design texts, and there is minimal or no discussion of research design in statistics texts. This book is intended to fill that gap.

I originally filled that gap with my own notes for the program evaluation class. For that class, I asked students to buy a research design text (usually the latest edition of Peter H. Rossi, Howard E. Freeman, and Mark W. Lipsey, *Evaluation: A Systematic Approach*, published by Sage Publications in successively updated editions since 1980). I also asked students to buy a statistics review text, often one or more of the "green books" in the Quantitative Applications in Social Science Series (also published by Sage). My notes (pages of typed weekly handouts) became the foundations of this book. The book is written for my students and for students like them. They are graduate students (and an occasional upperclassman) who have taken at least one statistics course and who are interested in public policy and public programs, not just from a normative standpoint but also

from the empirical standpoint of finding out what works. Specifically, this text is written not only for graduate students in professional public affairs master's degree programs (especially the MPP and other policy-oriented master's degrees) but also for students in other empirically oriented public affairs programs. This includes Ph.D. students in political science and sociology, students in applied master's degree programs in economics, and even some advanced undergraduates who have had a course in econometrics or statistical analysis.

This book is dedicated to the students who have taken my American University course, Public Program Evaluation. I asked them not only to learn how to describe, analyze, and critique the various research designs and corresponding statistical analyses that others used but also to complete an original program evaluation in one semester, with no outside funding. They always came through, although it was never clear in the middle of the course whether the project could be completed on time and within the zero-dollar budget. But we (that is, both the students and I) learned how to adapt the design to the data that were available and to apply the most appropriate statistical analysis for that design. After we collected the data, the fun began: crunching the numbers was exciting, we learned a lot during the process, and the end result was new knowledge, producing defensible information about a program that was not known before. Occasionally, the papers were even published in peer-reviewed journals.

I owe special thanks to students who took the time not only to read the draft chapters of this text but also to give me feedback. They were brave enough to tell me when I was unclear, misguided, or just plain wrong. They include Wei Song, Xiaodong Zhang, Reina Rusnak, and Pablo Sanabria. I owe the biggest thanks to my coauthor, Claire Felbinger. Without her supportive but continual nudges over the years, this book would not have been written. Of course, I am to blame for the remaining errors, obscurity, and lack of wisdom.

Since the original edition was written, Claire Felbinger died. This revision is dedicated to her memory, and to my fond memories of our many lunches where we discussed politics, performance measurement, and program evaluation.

I hope that you have as much fun systematically analyzing real data about real programs as my students and I have had.

Public Program Evaluation

This page intentionally left blank

1

What This Book Is About

WHAT IS PROGRAM EVALUATION?

Program evaluation is the application of empirical social science research methods to the process of judging the effectiveness of public policies, programs, or projects, as well as their management and implementation, for decision-making purposes. Although this definition appears straightforward, it contains many important points. Let us deconstruct the definition.

What We Evaluate: Policies, Programs, and Management

First, what is the difference between public programs, policies, and projects? For the purposes of this book, policies are the general rules set by governments that frame specific government-authorized programs or projects. Programs and projects implement policy. Programs are ongoing services or activities, while projects are one-time activities that are intended to have ongoing, long-term effects. In most cases, programs and projects, authorized by policies, are directed toward bringing about collectively shared ends. These ends, or goals, include the provision of social and other public services and the implementation of regulations designed to affect the behavior of individuals, businesses, or organizations.

While the examples in this text refer to government entities and public programs or projects, an important and growing use of program evaluation is by private organizations, including for- and not-for-profit institutions. Not-for-profit organizations frequently evaluate programs that they support or directly administer, and these programs may be partly, entirely, or not at all publicly funded. Similarly, for-profit organizations frequently operate and evaluate programs or projects that may (or may not) be funded by the government. Further, many organizations (public, not-for-profit, or for-profit) use program evaluation to assess the effectiveness of their internal management policies or programs. Thus, a more general definition of program evaluation might be the application of empirical social science research methods to the process of judging the effectiveness of organizational policies, programs, or projects, as well as their management and implementation, for decision-making purposes. I chose to focus on program evaluation in a public context because that is the predominant application of the field in the published literature. Private sector applications are less accessible because they are less likely to be published. Nonetheless, the research designs and other methods that this book elaborates are entirely portable from one sector to another.

The distinction between policy, program, and project may still not be clear. What looks like a policy to one person may look like a program or project to another. As an example, Temporary Assistance for Needy Families (TANF) can be regarded at the federal level as either a policy or

a program. From the state-level perspective, TANF is a national policy that frames fifty or more different state programs. Further, from the perspective of a local official, TANF is a state policy that frames numerous local programs. Similarly, from the perspective of a banker in Nairobi, structural adjustment is a policy of the International Monetary Fund (IMF). The loan, guaranteed by the IMF and serviced by the banker, funds a specific development program or project in Kenya, designed to carry out the market-building goals of structural adjustment policy. But from the perspective of an IMF official in Washington, DC, structural adjustment is just one of the programs that the IMF administers. So, whether one calls a particular government policy initiative or set of related activities a policy or a program (or project) depends on one's location within a multilevel system of governments. The distinction does not really matter for the purposes of this book. One could call this volume a book about either policy or program or even project evaluation. However, perhaps because the bulk of studies pertain to specific, ongoing, local programs, either in the United States or elsewhere, the common usage of the set of methods discussed in this book is to characterize them as program rather than policy evaluation. We regard projects as short-term programs.[1] (See Figure 1.1 for an example of the policy-program-project hierarchy.)

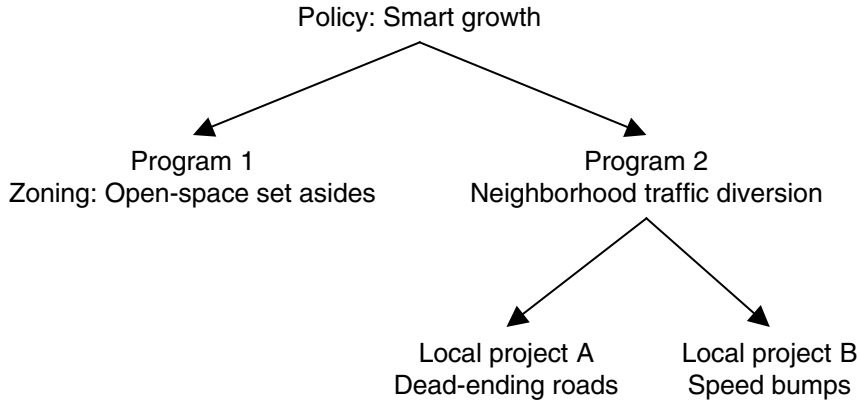
These examples also underscore another point: The methods in this book apply to government-authorized programs, whether they are administered in the United States or in other countries. While data collection may sometimes encounter greater obstacles in developing countries, this text does not focus on specific problems of (or opportunities for) evaluating specific programs in developing countries.[2] The logic of the general designs, however, is the same no matter where the research is to be done.

Finally, the logic of program evaluation also applies to the evaluation of program management and implementation. Evaluators study the effectiveness of programs. They also study the effectiveness of different management strategies (e.g., whether decentralized management of a specific program is more effective than centralized management) and evaluate the effectiveness of different implementation strategies. For example, they might compare flexible enforcement of a regulatory policy to rigid enforcement, or they might compare contracting out the delivery of services (say, of prisons or fire protection) to in-house provision of services.

The Importance of Being Empirical

As the first phrase of its definition indicates, program evaluation is an empirical enterprise. "Empirical" means that program evaluation is based on defensible observations. (I explain later what "defensible" means.) Program evaluation is not based on intuition. It is not based on norms or values held by the evaluators. It is based not on what the evaluators would prefer (because of norms or emotions) but on what they can defend observationally. For example, the preponderance of systematic empirical research (that is, research based on defensible observations) shows that gun control policies tend to reduce adult homicide.[3] An evaluator can oppose gun control policies, but not on the basis of the empirical claim that gun control policies do not reduce adult homicide. Evidence-based evaluation requires a high tolerance for ambiguity. For example, numerous studies, based on defensible observations show that policies allowing people to legally carry concealed weapons act as if they make it more risky for criminals to use their weapons because, empirically, these policies appear to reduce homicides rather than increase them. Equally defensible studies raise serious questions about the defensibility of that conclusion.[4] This example suggests that evaluation results based on defensible observations can sometimes produce ambiguous results. Most people's intuition is that it is dangerous to carry concealed weapons, but defensible observations appear to produce results that cause us to challenge and question our intuition.

Figure 1.1 **The Policy-Program-Project Hierarchy**



Although "empirical" refers to reliance on observations, not just any observations will do for program evaluation. Journalists rely on observations, but their observations are selected because of their interest, which may mean that they are atypical (or even typical) cases. Journalists may claim that the case is "atypical" or "typical" but have no systematic way to justify the empirical veracity of either claim. By contrast, program evaluation relies on the methods of science, which means that the data (the observations) are collected and analyzed in a carefully controlled manner. The controls may be experimental or statistical. In fact, the statistical and experimental controls that program evaluators use to collect and analyze their observations (i.e., their data) are the topic of this book. The book explains shows that, while some controls produce results that are more valid (i.e., defensible) than others, no single evaluation study is perfect and 100 percent valid. The book also sets forth the different kinds of validity and the trade-offs among the types of validity. For example, randomized experimental controls may rate highly on internal validity but may rate lower on external validity. The trade-offs among the different types of validity reinforce the claim that no study can be 100 percent valid.

The existence of trade-offs among different types of validity also makes it clear that observational claims about the effectiveness or ineffectiveness of public programs are not likely to be defensible until they are replicated by many studies, none of which is perfect. When many defensible studies, each imperfect in a different way, reach similar conclusions, we are likely to decide that the conclusion is defensible overall. We can act as if the conclusion is true. Note that we do not claim that the conclusion about the program is true, only that we can act as if it were true. Because no method of control can be 100 percent valid, no method of control leads to proof or truth. Mathematicians do proofs to analyze whether claims are true. Because program evaluators do not perform mathematical proofs, the words "prove" and "true" should not be part of their vocabulary.

Lenses for Evaluating Effectiveness

Program evaluation uses controlled observational methods to come to defensible conclusions about the effectiveness of public programs. Choosing the criteria for measuring effectiveness makes program evaluation not purely scientific; that choice is the value or normative part of the program evaluation enterprise. Programs can be evaluated according to many criteria of effectiveness. In

fact, each social science subdiscipline has different core values that it uses to evaluate programs. Program evaluation, as a multidisciplinary social science, is less wedded to any single normative value than most of the disciplinary social sciences, but it is important for program evaluators to recognize the connection between a criterion of effectiveness, the corresponding normative values, and the related social science subdiscipline.

For example, the core value in economics is allocative efficiency. According to this criterion, the increment in social (i.e., private plus external) benefits of a program should equal the increment in social (i.e., private plus external) costs of the program. This view leads economists to examine not only the intended effects of a program but also the unintended effects, which often include unanticipated and hidden costs. For example, the intended effect of seat belts and air bags is to save lives. An unintended effect is that, because seat belts and air bags (and driving a sport utility vehicle [SUV]) make drivers feel safer, protected drivers sometimes take more risks, thereby exacerbating the severity of damage to other vehicles in accidents and endangering the lives of pedestrians.[5] Thus, program evaluators, particularly those who are concerned with the ability of public programs to improve market efficiency, must estimate the impact of seat belts (or air bags or SUVs) not only on drivers' lives saved and drivers' accident cost reductions but on the loss of property or lives of other drivers or pedestrians.

A primary concern of economics that is also relevant to program evaluation is benefit-cost analysis, which ascribes dollar values both to program inputs and program outputs and outcomes, including intended as well as unintended, unanticipated, and hidden consequences. Program evaluation is critical to this exercise. In fact, program evaluation must actually precede the assignment of monetary values to program consequences, whether positive or negative. Specifically, before assigning a dollar value to program consequences, the evaluator must first ascertain that the result is attributable to the program and would not have occurred without it. For example, if program evaluation finds that using cell phones in automobiles had no effect on increasing accidents,[6] there would be no reason to place a dollar value on the increased accidents to measure an external cost of cell phones. That is, if a defensible conclusion is that cell phones in autos have no direct impact on increasing accidents, then the cost of the additional accidents cannot be ascribed to (or blamed on) cell phones in autos. In this example, the external cost of automobile cell phones in respect to increasing accidents would be zero. Program evaluation is critical to determine the beneficial (or harmful) effects of programs; it does not monetize the value of these effects.

Similarly, if a publicly funded job-training program fails to have an independent impact on improving employment outcomes for those in the training program, then it is clear that the program, which uses public resources, is not efficient. Any activity that uses resources and produces no valuable outcomes is clearly not efficient, although it might have other meritorious, noneconomic attributes. Moreover, ascribing consequences (whether positive or negative) to programs requires that the consequences can be attributed to the program, not to something else (e.g., an improving economy). This is a task for program evaluation, and it precedes benefit-cost analysis of the program.

Economists use program evaluation to assess program efficiency. Other social science subdisciplines have different core values and correspondingly different criteria for measuring effectiveness. Sociologists often focus on the core value of equality or social justice. Thus, a sociologist might evaluate welfare reform (the 1996 change from Aid to Families with Dependent Children [AFDC] to TANF) according to whether it lifts people out of poverty or reduces poverty gaps between blacks and whites. By contrast, an economist might evaluate welfare reform according to whether it makes people (including taxpayers and program beneficiaries) better off than they were before. Psychologists (and others) might examine how welfare reform

affected child welfare or parent–child relations. For program evaluation, all of these outcome criteria are relevant.

Political science and public administration, a close relation to political science, examine programs according to responsiveness or accountability. In political science, responsiveness might mean whether the program conforms to the preferences of voters or meets with the satisfaction of its clients. Political scientists might evaluate government services according to citizen or client satisfaction. For example, they might compare citizen or parent satisfaction in school districts that allow parents to choose schools with satisfaction in school districts that do not.[7]

Satisfaction is also important in program administration and management. For example, program managers often seek to satisfy clients, not only because satisfied clients might be one signal of program effectiveness but also because satisfaction helps to legitimize programs.[8]

Accountability examines whether the program is implemented in accordance with the preferences of the legislators who collectively authorize the program, appropriate money for it, write the detailed (or vague) rules for how it is to be implemented, and play a role in selecting the political appointee tasked with administering the program. As an example of accountability or bureaucratic responsiveness to political oversight, political scientists have found that aggressive implementation of environmental controls is more likely when political controls are in the hands of proenvironmental legislatures and political executives.[9]

Public administrators and program heads also have another core value: program effectiveness, gauged by empirical measures of program outputs and outcomes. Program outputs are usually short-term, countable results. Evaluations of program outputs may ask questions such as these: How many people are employed because of their participation in a training program? How much have achievement scores increased because of an education program or policy change? How many additional workplace inspections for compliance with environmental or workplace health and safety regulations were conducted as a result of a policy change? Does participation in school lunch programs increase school attendance?[10]

Compared to evaluations of program outputs, evaluations of program outcomes usually ask longer-range questions. Rather than examining the impact of job training on employment, outcome evaluations might assess long-run earnings. Similarly, rather than examine the impact of a school reform on achievement scores, an outcome evaluation may examine whether the reform resulted in higher graduation rates or better jobs. Or evaluators might examine whether more inspection leads to less pollution or enhanced workplace safety or if privatizing solid waste collection leads to long-run cost savings.[11]

The basic point is that program evaluation uses multiple lenses to assess program performance. These include the norms of economic efficiency, equality and social justice, responsiveness and accountability, and aspects of effectiveness measured by short-term outputs or long-term outcomes. For program evaluation, the general rule is to be attentive to multiple values and multiple perspectives. This means that program evaluators should use multiple outcome indicators that correspond to multiple values in judging program performance. Evaluators must also pay particular attention to the values and perspectives held by important stakeholders in the program. Stakeholders include important political decision makers, interested groups and individuals, program managers, program clients, and perhaps others as well. Their input helps to ensure that program evaluations examine multiple goals that are important to at least one of the affected parties. The result is that evaluators usually conclude that programs are effective and perform well in one dimension but not in another.[12] Another result is that evaluations may engender or contribute to political debate, improving the ability of legislators to do their job of representation.[13]

Providing Deliverables

A final point in deconstructing the definition of program evaluation is that it is used "for decision-making purposes." This point sets program evaluation apart from academic research. Academic social science research may meet all of the foregoing criteria because it is empirical research based on defensible observations and uses multiple evaluation criteria.[14] But, unlike research conducted for program evaluation, academic research does not need to be practical, on time, and clear to non-academics. The primary audience is likely to be other academics in the author's field. Academic research will not be published until it has undergone blind review by peers, who have an incentive to look for reasons not to publish someone else's research (but to publish their own). Peer review is time-consuming. It is not uncommon for five years to elapse between the time the data are collected and analyzed and the publication of study results. Before publication, many studies undergo reanalysis and revision during multiple peer reviews. Under this system, while academics labor under budget constraints, they do not face severe time constraints, especially if they are tenured.

In contrast, program evaluation information, which usually does not undergo blind review by peers, is most useful only when decisions have to be made. If the legislative vote on program reauthorization is next Thursday, then the information that arrives on the following Friday is no longer useful, at least for affecting the vote. If the program manager's budget request is due on September 1, information that arrives on September 8 is not very useful. If the program evaluation report is framed in theoretical terms, then it is less likely to be attended to by program administrators and other decision makers, though it is more likely to be published in an academic journal. For example, if an evaluator were to frame a study of the impact of performance pay on teacher effectiveness in terms of theories on optimal incentive systems, incomplete contracts, adverse selection, moral hazard, and rational choice, decision makers might ignore it. If the same study is framed less abstractly, decision makers will be more likely to pay attention to its findings. (There is a large and growing academic literature on the rationale for and impact of linking teacher pay to student performance. There is less evidence that the literature actually influences the choice of pay systems for teachers.[15])

Further, the information must be clear. Academics can talk about unbiased regression parameter estimates, *t*-scores, *p*-values, homoscedasticity, and the like, but program evaluators have to explain what this shorthand jargon means. Further, brevity is an important component of clarity: for program evaluators, a crisp 300-word abstract or executive summary or a few (less than five) PowerPoint® slides (backed by a more detailed report) is critical for communication to busy decision makers.[16]

In addition to timeliness and clarity, program evaluation faces another constraint that academic research may not confront. Because its primary consumers are decision makers and not other researchers, those who demand (and authorize the payment for) program evaluation may not value the intricacies of defensible research as much as evaluators themselves do. The result is that decision makers are often reluctant to support adequate budgets for evaluation research. Evaluators thus face not only time constraints but also budget constraints that shape the research designs that they use. The most valid design from an academic perspective may be infeasible for the program evaluator, who often faces more binding money and time constraints.

TYPES OF PROGRAM EVALUATIONS

While program evaluations share a common definition, there are several kinds of program evaluation studies. Each asks a somewhat different question. There are four basic types of program

evaluation questions. Each type can be classified in terms of its methodology and in terms of its substantive focus. In terms of method, program evaluations can be largely descriptive or they can be focused on assessing causal claims. In terms of substance, evaluations can focus on program inputs or outputs and outcomes.

Focus on Method: Descriptive and Causal Evaluations

Descriptive methods are essentially "bean counting." They are used to describe a population often (but not always) based on a random sample.[17] Descriptive evaluations might estimate the characteristics of a target population of potential clients. For example, if the target population is preschool children in families below the poverty level, a descriptive evaluation might assess their nutritional status, their overall health, their verbal and numerical abilities, and so on. "Needs assessments" are a common type of descriptive program evaluation that describes a target population of potential clients. Another type of target population is the population not of clients but of existing programs. For example, this type of descriptive evaluation might assess the characteristics of the population of local recycling programs. Such an evaluation might estimate their average budget, describe who administers the program (e.g., the government itself or a contractor), characterize the type of program (e.g., curbside pickup, payment for bottle return, or per-pound charge for curbside waste), or describe the types and tons of waste recycled. Thus, descriptive programs may assess program inputs, outputs, or both.

Another type of program evaluation is causal. As the term implies, causal evaluations aim to assess cause: Did the program bring about desired outputs and outcomes and avoid undesired outputs and outcomes? This is equivalent to asking the question in terms of cause: Did the program cause the desired outcomes (without causing undesired ones)? Another way to ask the same question is to ask whether the output or outcome would have happened anyway. For example, if clients would have found above-minimum-wage employment even without the job-training program, then it would be difficult to credit the job-training program with causing the outcome. Conversely, if a particular extent of school segregation would occur even without a program to bus children to magnet schools, it would be difficult to blame the program for causing the adverse outcome.

Substantive Focus: Implementation and Output and Outcome Evaluations

In terms of the substantive focus of the program evaluation, some evaluations center on the implementation of programs, while other evaluations focus on the outputs or outcomes of programs.

Examples of evaluations that examine implementation include assessments of program accountability and program management that are common in public administration and political science. They are often referred to as formative evaluations. For instance, studies of whether programs are implemented "as planned" or "on time" and "on budget" fall into the class of implementation or formative evaluations. Similarly, studies of program management also fall into this class of process or implementation evaluations. These studies might examine whether the program services are contracted out or run by the agency; whether the program activities are organized functionally or geographically; whether programs are administered centrally or locally; or whether the program staff is paid based on individual or program performance or face flat pay or a mix of flat and variable pay.

By contrast, evaluations that (exclusively) examine outputs or outcomes are most common in psychology, economics, and sociology. They are often referred to as summative evaluations. For instance, common output measures in assessments of education programs are standardized

test scores, while outcome measures might include graduation rates or postgraduation measures, such as employment status or wages and salary. Output measures that are commonly used in the assessment of welfare and job-training programs include number of clients served and number of training hours per client; outcome measures might include clients' earnings, poverty levels, job mobility, and long-term wage increases. Regulatory program outputs include making inspections, assessing penalties, and other indicators of compliance or noncompliance. Regulatory program outcomes include measures of pollution, safety, and health.

Four Types of Evaluations

The two classes of methodological evaluations (descriptive vs. causal) and the two classes of substantive lenses (implementation vs. output/outcome) can be cross-classified as four basic types of program evaluations, as shown in Table 1.1. The types of questions that these evaluations ask are listed in each box.[18]

Using descriptive methods and focusing on program implementation, one type of program evaluation focuses on how the program was implemented. Questions include: What activities does the program support? What is the purpose or goal of these activities? Who performs these activities? How extensive and costly are the activities, and whom do they reach? Are conditions, activities, purposes, and clients fairly similar throughout the program, or is there substantial variation across program components, providers, or subgroups of clients? Were these the intended activities, costs, procedures, and time lines?

Using descriptive methods and focusing on program outputs and outcomes, another type of program evaluation is targeted at the objects at which a program's activities are directed and the coverage that these activities achieve. Questions include: Have program efforts focused on the intended targets or problems? Has the program reached the appropriate or intended people or organizations? Do current resources and targeting practices leave significant needs unmet? Assessing a program's targeting success typically requires not only data on the clients served but also information about the size and distribution of the eligible population, target group, or problem area as a whole.

Using causal methods and focusing on program outputs and outcomes, the third type of program evaluations estimates the impacts of public programs. Questions include: Did the program achieve its intended purposes or outcomes (impact), or would the outcome have happened anyway? Did the program have unintended beneficial or adverse effects? How does the current program compare to alternative strategies for achieving the same end (comparative advantage)? These impact questions center on whether program activities actually resulted in the improvements that the program was designed to produce. That is, the focus is on whether program activities are consistent with the hypothesis that they actually caused these improvements (or adverse effects). For example, a job-training program is expected to show that participation in the program led to significantly higher earnings or more stable employment. Similarly, regulatory programs are expected to show that implementation leads to reduction in the targeted hazard and that the reduction would not have happened otherwise.

Causal evaluations do not always appear in the guise of the explicit language of causality. That is why I frame the first two illustrative causal/output/outcome questions in Table 1.1 in two ways. One version uses the explicit language of causation (Does the use of seat belts cause fewer deaths? Do smaller classes cause higher achievement?). The other version is equivalent, but the word "cause" does not appear in these forms of the question, so the issue of causation is implicit but still central (Do seat belts save lives? Do smaller classes raise achievement?). The absence of the word "cause" does not signify the absence of the causal issue. (See Table 1.1.)

Table 1.1

Four Basic Types of Program Evaluations

Methodological focus	Substantive focus: Implementation	Substantive focus: Output/Outcome
	Illustrative types of questions	
Descriptive	How many are enrolled? How many classes are offered? How many inspections were carried out? What is the ratio of servers to clients? What are the costs per client? Does the program operate • on budget? • on time? • in compliance with rules? What percentage of eligibles is served? Of those served, what percent is ineligible? Are clients in high-poverty areas (or in different jurisdictions) served more or less than those in other areas?	How many/what percent of clients are/is employed/above poverty/in school? What is the number/rate of workplace: • accidents/toxic substances observed? • pollutants emitted? What are scores on standardized tests? How do the above measures vary: • by jurisdiction? • by program need? What is the overall crime rate? • property crime rate? • violent crime rate? How do the crime rates vary: • by jurisdiction? • by poverty rate?
Causal	Why do (what causes) some jurisdictions (to): • implement regulations more strictly? • delay implementation? • provide more generous/extensive benefits/services? • offer a different mix of services? Does management affect implementation? Does contracting out save money/speed implementation? Does decentralized management: • reduce case worker discretion? • improve program performance?	Do seat belts save lives? (i.e., Do seat belts cause fewer deaths?) Do smaller class sizes raise achievement? (i.e., Do smaller classes cause higher achievement?) Do computers in classrooms raise student achievement? Do stricter regulations: • increase compliance? • reduce illness/pollution/accidents? Does training/education increase probability/quality of employment? Does increasing/targeting police patrol reduce crime rates? Does decentralized management result in better client outcomes?

A third version of a causal question about outcomes appears as seeking an answer to "Why?" For example, if one asks why some people are homeless, one is asking an implicit causal question: What causes some people (and not others) to be homeless? Program evaluators may ask questions in this form because the answer may cast light on program impacts. For example, evaluators may find that among the causes of homelessness are various policies that restrict or reduce the supply of low-income housing, such as zoning laws, restrictive rent control policies, and community development activities. Thus, causal questions about outcomes (or outputs) appear in various guises:

Does program *X* cause outcome *Y*?
Does program *X* have outcome *Y* (i.e., does *X* cause *Y* to occur?)?
What causes outcome *Y* to vary (among persons or places or over time)?
Why does outcome *Y* vary (i.e., what causes *Y* to vary?)?

Using causal methods and focusing on program implementation, the fourth type of program evaluation pays more attention to "why." It asks: Why are programs implemented differently? Why are some states more generous with welfare programs than others? Why do state taxes vary? Why do some cities contract out refuse collection (or other services) and others do not? Why are some school districts more supportive of school choice initiatives than others? But the underlying issue is still a causal question about program implementation, management, or administration: What causes some jurisdictions to be more aggressive or generous or speedy at implementing a program than others?[19]

Hybrid Evaluations

It is important to note that the 2 × 2 classification in Table 1.1 omits an important class of causal program evaluation questions. The omitted class represents the intersection of causation, implementation, and outcome. These evaluations explicitly link implementation to outcome, asking whether a characteristic of the implementation *caused* the outcome. The causal link is critical to these evaluations because, if the outcome would have occurred anyway, the causal claim about the method or mode of implementation cannot be sustained.[20] A typical question in this hybrid class of evaluation questions might be whether smaller classes (a question of implementation) cause higher achievement (an outcome).[21] The Student Teacher Achievement Ratio (STAR) experiment carried out in Tennessee in the 1980s suggests that smaller class size improves student test scores in the lower primary grades.[22] Another hybrid evaluation question asks whether increasing levels of implementation of a regulation cause higher compliance (an intended output) and lower pollution or better health (intended outcomes). Some studies have shown that additional health inspections, or inspections imposing penalties, decrease workplace injuries,[23] while other studies show that penalties imposed by the Occupational Safety and Health Administration (OSHA) are too small to have any significant deterrence incentive for employers to reduce workplace injury.[24] Studies that compare the impact of private, public, and nonprofit service delivery (e.g., trash collection, hospitals, K–12 schools, job training, prisons) on costs, outputs, or outcomes are also examples of hybrid evaluations.

Causal Evaluations

This book centers on the general class of causal questions because they are the hardest to answer. Although all the criteria for assessing the validity (i.e., defensibility) of descriptive claims are also used to assess the validity of causal claims, the reverse is not true. Specifically, as Chapter 2 explains, there are four general criteria for validity: external, measurement, statistical, and internal validity. Causal claims are unique because they have to meet one validity criterion, internal validity, which is not applicable to descriptive claims. All the other validity criteria apply to both causal and descriptive evaluations. Thus, while this book concentrates on methods for improving the defensibility or validity of causal claims about program impact, it will also apply, by logical extension, to methods for assessing the validity of descriptive claims about public programs.

In practice, most observational program evaluations fall into more than one category of the typology drawn above. For example, virtually all evaluations of program impact describe the "typical" program in the sample that is being studied and the variation from, say, one region to another or the variation in the size of the programs that are being studied. Similarly, most researchers who study why programs are implemented differently in various jurisdictions also describe variations in the range of programs under study. And, as noted earlier, some evaluations are hybrids.

Hybrid evaluations not only fall into both classes of causal evaluations but often also include descriptive components. These evaluations examine why programs are implemented differently in different jurisdictions (causal/implementation) and the impact of different implementation strategies on program outputs or outcomes (causal/output or outcome). They also describe the level and variation in both implementation and outputs or outcomes (descriptive implementation and output/outcome). Thus, evaluations like these fall simultaneously into all four categories displayed in Table 1.1. For example, Scholz and Wei describe how state and local governments implement OSHA workplace safety regulations and levels of firm compliance.[25] They also examine why OSHA workplace regulations are implemented more rigidly in some areas than others. That is, they ask: Why are enforcement activities (e.g., inspections) more frequent and penalties more severe in some jurisdictions than others, even when the characteristics of the firms are similar? Thus, they are asking what causes variation in implementation strategies. The same study then goes on to examine whether more rigid enforcement causes increases in levels of compliance.[26]

BASIC CHARACTERISTICS OF PROGRAM EVALUATION

Retrospection

This is a book about empirical, or observationally based, program evaluation. Program evaluations like these, whether descriptive or causal, are retrospective or contemporaneous. That is, evaluators cannot use observational data to describe or assess the actual impact of programs that do not (or did not) exist. Decision makers may use evaluators' assessments about previous or ongoing programs to make conjectures about the program's future characteristics or effects, but the core of program evaluation is a retrospective or contemporaneous assessment of previous or ongoing programs rather than the evaluation of future programs. Evaluators may assess imaginary programs by establishing a trial program and assessing it. However, after the trial program is established, it becomes an ongoing "trial," and it can be assessed as a contemporaneous evaluation.

Comparison

This is also a book about causal program evaluations. The primary purpose of causal program evaluations is to establish defensible causal claims about program implementation and impact. *It is impossible to make causal claims in the absence of comparison.* Two types of comparison lie behind any observationally based causal claim. One type compares entities with the program to those without it, or it compares entities with a lot of the program to those with just a little. Typically, this kind of comparison is called a cross-section (CS) comparison, because two (and often more than two) different entities (with and without) are being compared at the same time. For example, one might compare the academic performance (or the delinquency record) of high school students who participate in competitive sports programs with that of (comparable) students who do not—this is a with-and-without CS comparison. Or one might compare the achievement levels of children whose mothers face fairly lenient TANF work requirements with the achievement levels of similar children whose mothers face stricter or more time-consuming weekly TANF work requirements. This is an "a lot, a little" CS comparison. Or one might compare stakeholder satisfaction with conventionally written administrative regulations to stakeholder satisfaction with (comparable) negotiated administrative rules.[27] This is a "Type A vs. Type B" CS comparison. There are many kinds of CS comparison designs seen in the chapters that follow on experimental, quasi-, and nonexperimental designs.

Another type of comparison is before-and-after, or a time-series (TS), comparison. In this case, one compares the same entity (or set of entities) over time, examining how the entity operated before being subject to the program and then examining the same entity afterward. For example, one can examine whether raising state speed limits from 55 to 65 miles per hour (mph) increases accidents by comparing accident rates in a state before and after the speed limit was raised. Similarly, one might compare smoking rates in states for a period before and after cigarette taxes were raised. TS evaluations like these are particularly useful for "full coverage" programs. In this case, there is no entity that is not affected by the program, making a with-and-without CS comparison impossible. I discuss TS comparisons in the chapters that follow on quasi- and non-experimental designs. A TS-only comparison never applies to true experimental designs.

Sometimes evaluators make both kinds of comparisons at the same time. For example, some states have raised speed limits to 65 mph, while others have not. Some states have implemented TANF work requirements more strictly than others, and they implement (and change) these requirements at different points in time. Some high schools within a single school district adopt work-study programs, while others do not. In these cases, one can simultaneously make before-and-after (TS) and with-and-without (CS) comparisons. For instance, an evaluator can compare accident rates in states before and after each one raised its speed limit from 55 to 65 mph, simultaneously comparing accident rates in states with the lower limit to the rates in similar states with the higher limit. Or one can compare employment rates and salaries within a state before and after state TANF implementation and between states with different TANF implementation. One could compare alcohol sales between states with and without a state monopoly on retail alcohol to sales and simultaneously examine what happens to sales within states that change from monopoly to none.[28] I discuss these mixed TS and CS designs in the forthcoming chapters on quasi- and non-experimental evaluation designs.

In any case, some kind of comparison is necessary for causal evaluation. Causal evaluations with no comparison (that is, with no counterfactual or net effect) are not defensible.[29] Comparison is necessary (but not sufficient) for defensible causal claims. Examining one entity with no comparison at all makes it impossible to answer whether having the program (or not) makes (i.e., causes) a difference. When there is no comparison, the researcher can examine one entity and describe both the program and the level (and the distribution) of the output/outcome variables after the program was implemented. The researcher can even compare the level of the outcome variables to an external standard. This is often done in a performance evaluation, such as those required by the Government Performance and Results Act (GPRA) or the Program Assessment and Rating Tool (PART). But the researcher cannot defensibly conclude that the program caused the outcome because, without some minimal comparison, one cannot rule out the possibility that the level of the outcome would have been the same whether the program was implemented or not.

As an example, an evaluator can describe the elements of a high school work-study program and the achievement levels, delinquency rates, and part-time employment of students in the work-study program after it has been implemented. But without comparison (e.g., to outcomes of the students before participation or to outcomes of similar nonparticipants), no causal claim about the impact of the work-study program on the performance of the students is possible. Note that if the evaluator observed that achievement scores in the target group went up after the program was implemented, she is making a TS comparison. There would also be comparison if she observed that students in the school with the work-study program have higher achievement scores than similar students in the nearby school without the program. This would illustrate a CS comparison. But both of these examples comprise comparison.

Causal claims with no comparison at all are clearly indefensible. Making defensible causal claims about program impact even when there is comparison is not easy. In fact, it is the topic

of this book. We will see that some types of comparison are more defensible than others. That is, different types of research designs make different kinds of comparisons, and some of these designs are more defensible for making causal claims. Unfortunately, the designs that are the most defensible in this regard are often less defensible in other ways.

RELATION OF PROGRAM EVALUATION TO THE GENERAL FIELD OF POLICY ANALYSIS

The general field of policy analysis is much broader than program evaluation, which can be regarded as a subfield of public policy analysis. Both fields are explicitly multidisciplinary, drawing on the applied portions of the social and natural sciences. Thus, neither policy analysis nor program evaluation is purely theoretical. Both fields include practitioners with backgrounds from many of the (applied) social and natural sciences. So, program evaluators (or policy analysts) may have degrees in program evaluation (or public policy), but many have degrees in psychology, economics, sociology, political science, chemistry, biology, ecology, and so on. Further, as I indicated earlier, the pure social sciences are not only concerned with abstract theories; they also have core normative values that guide the central theoretical questions of each social science discipline. By contrast, program evaluation, as a general field of study, has no core substantive norm. That is not to say that the field lacks core values. Rather, the core value is using defensible, observationally based (i.e., empirical) evidence on program performance for informing decisions about the termination, modification, continuation, or expansion of public programs. Its core value is allegiance to the scientific method and to its ethical use in applied empirical policy research, not to particular substantive normative values or to disciplinary theory. Rather, it uses multiple outcome criteria in the context of real-world programs and program options.[30] While I argue that program evaluation ought to pay more attention to disciplinary theories (from social and natural science) to frame evaluators' expectations about likely policy or program outcomes, the primary focus of the field is applied, defensible, empirical research about policy or management programs that can be used to inform collective decisions. Its primary task is to provide an answer to the question "What works?" that is as valid as possible.[31]

ASSESSING GOVERNMENT PERFORMANCE: PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

Program evaluation is not the same as performance measurement. Performance measures are the indicators of outputs or outcomes (or implementation or management strategies) that are used in program evaluation. Program evaluation requires performance measures but entails other demands as well, including multiple types of validity. Selecting the indicators that are used for performance measurement, or in a program evaluation, is a normative and political choice. It is also a choice that can affect how a program operates. As an example, consider the performance measures used for the Workforce Improvement Act (WIA) of 1998, which replaced the Job Training Partnership Act. It requires states to establish measurable performance standards, such as the following:

- the number of adults who obtained jobs by the end of the first quarter after program exit, excluding those employed at registration
- the number of adults with jobs in the first quarter after program exit, and the number with jobs in the third quarter after exit
- the number of adults with jobs in the first quarter after program exit, and the increases in their post-program earnings relative to pre-program earnings.

If program managers want to appear "successful," then they will select for training the least needy of those who are eligible (known as "creaming").[32]

The federal government uses two different systems of performance measurement. (State and local governments use similar systems to assess performance.) The first is the GPRA. Passed by Congress in 1993, it requires all agencies to submit to the Office of Management and Budget (OMB) an annual performance plan covering each program activity set forth in the agency's budget. These plans are to be consistent with the agency's strategic plan and should include the following features:

- the establishment of goals to define the level of performance to be achieved by a given program activity
- the use of goal statements that are objective, quantifiable, and measurable
- the use of performance indicators to measure or assess the relevant outputs, outcomes, or service levels for each program activity
- a description of the operational processes and resources required to meet the performance goals.

It should be apparent from this list that the core of the GPRA is descriptive program evaluation, spanning the description of both outcomes and implementation, and assessing them against externally set goals or standards. Although the GPRA, as applied in some agencies, may require the analysis of whether the program actually causes the outcomes, impact analysis is not the core of the GPRA. For example, under the GPRA, the Internal Revenue Service (IRS) might set, as one measure of compliance, a standard that 95 percent of all individual and corporate tax returns that are supposed to be filed by April 15 will be filed by that date. If that outcome is empirically observed, one cannot conclude that it was accomplished because of any activities or policies or programs of the IRS. In fact, it is even possible that, had the IRS done something else (e.g., changed the penalties for late filing), the compliance rate might have been even higher. Similarly, the WIA performance measures listed previously are outcome measures that might be used in a descriptive program evaluation. Alone, however, these measures cannot establish whether the WIA program, or something else, brought about the outcome. For example, whether adults obtain jobs or increase their earnings may also be attributable to changes in the local or state economy or in the local labor supply.

The executive branch uses a more recent, competing system of performance measurement to manage disparate federal agencies.[33] The PART is administered by the OMB as part of presidential oversight of federal agencies. The first ratings appeared in 2004; by now, nearly all federal programs have PART scores, and many programs have received PART scores in multiple years, to measure change. PART scores are used to assess the following four features of programs:

- clarity of program purpose and whether it is well designed to achieve its purposes
- strategic planning: Does the program have "valid" annual and long-term goals?
- program management, including financial oversight and program improvement efforts
- results that programs can report with "accuracy" and "consistency."

Performance measurement is necessary for program evaluation. Specifically, descriptive program evaluations require valid measures of performance. Forthcoming chapters define validity and describe how to assess it. Causal evaluations, however, start from measuring and observing