Contents

List of Figures and Tables                                       ix

Preface and Acknowledgments                                      xiii

Introduction                                                     xvii

Part I               Some Underlying Ideas                          1

Chapter 1       Evaluation in Perspective                           3

   Relating Evaluation to Purposeful Management
      Behavior                                                     3
   Exploring Organizational Boundaries                            7
   Evaluation                                                     8
   Direct Intervention                                            10
   Those In Charge                                                13
   What, then, Is an Evaluator to Do?                             16

Chapter 2       Expanding upon Evaluation as a Part of Purposeful
                  Behavior                                                  19

   Operation of a Simple, Mechanistic, Feedback System            21
   Why is Life like a Furnace                                     27

Part II             Describing the Universe: Models and Measurement   33

Chapter 3       Models                                              19

   Representing Complex Social Systems and
      Organizations                                                 35
   Finding the Direct Intervention: Where Functional
      Models Begin                                                   36
   Drawing an Example from the Home Heating System                 38
   Why Make a Model of the Direct Intervention?                   53

Chapter 4       Measurement                                         59

   How High Is this Table from the Floor?: Steps in
      Measurement                                                   60
   Some Kinds of Error                                             66
# Government Oversight and Evaluability Assessment

## Part III

**The Domain of Those in Charge**

<table>
<thead>
<tr>
<th>Chapter 5</th>
<th>Those in Charge of Government Agencies</th>
<th>75</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The Planners’ Dream of the Government Organization</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>The Government Organization during Waking Hours</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>Multiple Levels of Organization</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>Multiple Levels of Semantics</td>
<td>81</td>
</tr>
</tbody>
</table>

## Chapter 6

**What is Acceptable to Those in Charge**

| Why is Acceptability a Problem? | 87 |
| Factors Influencing Acceptability | 91 |
| Analytic Integrity and Political Efficacy | 96 |

## Chapter 7

**Some Clues on Exploring in the Those-In-Charge Domain**

| The Treasure Hunt | 99 |

## Chapter 8

**Purposeful Behavior Revisited: First Find out What the Purpose Really Is**

| Where Organizational Incentives Come From | 109 |
| Effects of Organizational Incentives on the Individual | 110 |
| What to Look For | 113 |
| Why Appropriate Measures and Comparisons May Meet either Acceptance or Resistance | 117 |

## Part IV

**Evaluability Assessment or We Said all That to Say This**

| Chapter 9 | Evaluability Assessment: An Overview | 123 |
| Testable and Equivalency Information | 125 |
| Comparisons and Reconciliations within and between Families of Models | 126 |
| Sequential Purchase of Information | 129 |
| Two Ways to Foul-Up an Attempt at Oversight | 133 |
| Big Organizations Make It All Harder: Hierarchies and Discontinuities | 134 |
| Synopsis | 139 |
| | 140 |
Chapter 10  Testable Information  143

The Testable Logic Model  144
The Testable Functional Model  149
Discussion and Validation  151
Warning  155

Chapter 11  Equivalency Information  159

The Preliminaries  159
Building the Equivalency Models  160
Alarums and Excursions  166
Level of Detail  167
Convergence  168
The Hybrid Testable-Equivalency Model  170

Chapter 12  When is a Program Evaluable  175

Evaluable Programs  176
Comparisons and Test of Evaluability  180
Summary  193

Chapter 13  The Sequential Purchase of Information  195

Some Examples of Why Costs and Effort Vary  200
Acceptance of Sequential Purchase  206

Chapter 14  Sources of Information for EA Models  207

Sources of Information and the Related Models  208
Recapitulation and Coda  219

Chapter 15  Afterword  223

Appendix  The Creation and Development of EA  225

Index  235

About the Authors  243
# List of Figures and Tables

## Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>Evaluation in Perspective</td>
<td>4</td>
</tr>
<tr>
<td>1-2</td>
<td>A Paradigm of Purposeful Behavior with 3 Domains</td>
<td>6</td>
</tr>
<tr>
<td>1-3</td>
<td>A Simple Paradigm</td>
<td>8</td>
</tr>
<tr>
<td>1-4</td>
<td>Laying Out and Measuring a Direct Intervention</td>
<td>11</td>
</tr>
<tr>
<td>1-5</td>
<td>Those in Charge</td>
<td>15</td>
</tr>
<tr>
<td>2-1</td>
<td>A Home Heating System</td>
<td>22</td>
</tr>
<tr>
<td>2-2</td>
<td>A Home Furnace with Policy Management Included</td>
<td>25</td>
</tr>
<tr>
<td>2-3</td>
<td>A Social Program</td>
<td>30</td>
</tr>
<tr>
<td>3-1</td>
<td>The Mystery Model</td>
<td>39</td>
</tr>
<tr>
<td>3-2</td>
<td>A Rudimentary Model</td>
<td>40</td>
</tr>
<tr>
<td>3-3</td>
<td>The Furnace Company and the House as Black Boxes</td>
<td>41</td>
</tr>
<tr>
<td>3-4</td>
<td>Logic Model of the Office Personnel</td>
<td>42</td>
</tr>
<tr>
<td>3-5</td>
<td>Two Models of a Home Heating System</td>
<td>42</td>
</tr>
<tr>
<td>3-6</td>
<td>Functional Model Arranged to Trace Energy Flow (Control Circuits Omitted)</td>
<td>44</td>
</tr>
<tr>
<td>3-7</td>
<td>Five Levels of Models of a Heating System</td>
<td>46</td>
</tr>
<tr>
<td>3-8</td>
<td>Garbage Transfer</td>
<td>48</td>
</tr>
<tr>
<td>3-9</td>
<td>Mystery Model of a School System</td>
<td>51</td>
</tr>
<tr>
<td>3-10</td>
<td>Functional Model Arranged to Trace Knowledge Transfer (Teaching and Technique Knowledge Omitted)</td>
<td>52</td>
</tr>
<tr>
<td>3-11</td>
<td>Knowledge Transfer Model</td>
<td>53</td>
</tr>
<tr>
<td>3-12</td>
<td>Three Types of Models: Logic, Measurement and Functional</td>
<td>56</td>
</tr>
<tr>
<td>5-1</td>
<td>The Planners' Dream of a Governmental Organization</td>
<td>77</td>
</tr>
<tr>
<td>5-2</td>
<td>Levels in Organization</td>
<td>80</td>
</tr>
<tr>
<td>6-1</td>
<td>Examples of Some Possible Measures and Comparisons</td>
<td>89</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>8-1</td>
<td>The Individual Comparing the Outcome of Action with the Local Market Standard</td>
<td>115</td>
</tr>
<tr>
<td>8-2</td>
<td>The Individual Comparing the Outcome of Action with Local and Exogenous Market Standards and Rank Ordering the Importance of the Markets</td>
<td>116</td>
</tr>
<tr>
<td>9-1</td>
<td>Basic Distinctions Used in Evaluability Assessment</td>
<td>127</td>
</tr>
<tr>
<td>9-2</td>
<td>Testable Functional Model of Child-Health Projects</td>
<td>128</td>
</tr>
<tr>
<td>9-3</td>
<td>Equivalency Functional Model of Child-Health Project I</td>
<td>130</td>
</tr>
<tr>
<td>9-4</td>
<td>Equivalency Functional Model of Child-Health Project II</td>
<td>131</td>
</tr>
<tr>
<td>10-1</td>
<td>One Overall Logic Model for Universal, Free Education</td>
<td>144</td>
</tr>
<tr>
<td>10-2</td>
<td>Superintendent's Logic Model for CAI Program</td>
<td>145</td>
</tr>
<tr>
<td>10-3</td>
<td>Assistant Superintendent's Logic Model for CAI Program</td>
<td>145</td>
</tr>
<tr>
<td>10-4</td>
<td>Curriculum Coordinator's Logic Model for CAI Program</td>
<td>146</td>
</tr>
<tr>
<td>10-5</td>
<td>Director of Evaluation's Logic Model for CAI Program</td>
<td>146</td>
</tr>
<tr>
<td>10-6</td>
<td>Principal's Logic Model for CAI Program</td>
<td>147</td>
</tr>
<tr>
<td>10-7</td>
<td>Example of a Combined-Logic Model for CAI Program</td>
<td>148</td>
</tr>
<tr>
<td>10-8</td>
<td>Laying Out a Direct Intervention</td>
<td>150</td>
</tr>
<tr>
<td>10-9</td>
<td>Black-Box Testable Functional Model of CAI Program</td>
<td>150</td>
</tr>
<tr>
<td>10-10</td>
<td>Some Elements of a Testable Functional Model of the CAI Direct-Intervention and Comparison Groups</td>
<td>152</td>
</tr>
<tr>
<td>10-11</td>
<td>Additional Detail on the Instruction Testable Functional Model</td>
<td>153</td>
</tr>
<tr>
<td>10-12</td>
<td>Some Suggested Measurements Based on Testable Functional Model of CAI Program</td>
<td>156</td>
</tr>
<tr>
<td>11-1</td>
<td>Black-Box Equivalency Functional Model of Actual Program</td>
<td>162</td>
</tr>
<tr>
<td>Figure/Table</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>11-2</td>
<td>Some Elements of an Equivalency Functional Model of the Textbook-Demonstration-Program Direct Intervention</td>
<td>164</td>
</tr>
<tr>
<td>11-3</td>
<td>Additional Detail on the Textbook-Demonstration Equivalency Functional Model</td>
<td>165</td>
</tr>
<tr>
<td>11-4</td>
<td>A Hybrid Testable Equivalency Model of the Textbook Demonstration</td>
<td>172</td>
</tr>
<tr>
<td>12-1</td>
<td>What is Evaluable? Two Kinds of Comparisons</td>
<td>178</td>
</tr>
<tr>
<td>12-2</td>
<td>A Basic Format for Laying Out Evaluable Questions and the Means for Answering Them</td>
<td>181</td>
</tr>
<tr>
<td>12-3</td>
<td>Some Elements of a Hybrid Functional Model of the Textbook Demonstration</td>
<td>182</td>
</tr>
<tr>
<td>12-4</td>
<td>Model, Measures, and Issues Related</td>
<td>183</td>
</tr>
<tr>
<td>12-5</td>
<td>Some Evaluable Questions for a Child-Health Project</td>
<td>186</td>
</tr>
<tr>
<td>12-6</td>
<td>Logic Underlying Section 111(jj) and 1133(d)(4) of the 1977 Clean Air Act Amendments</td>
<td>187</td>
</tr>
<tr>
<td>12-7</td>
<td>Hybrid Functional Model of Homer City Plants and the Direct Intervention: An Innovation in Coal Cleaning</td>
<td>189</td>
</tr>
<tr>
<td>12-8</td>
<td>Sample Elements of a Model with Multiple Domains: The Louisville Emission-Offset Bank</td>
<td>191</td>
</tr>
<tr>
<td>12-9</td>
<td>Evaluable Model of an Offset Bank</td>
<td>192</td>
</tr>
<tr>
<td>13-1</td>
<td>A Paradigm for Sequential Purchase</td>
<td>196</td>
</tr>
<tr>
<td>13-2</td>
<td>Simplest Hazardous-Waste Flow</td>
<td>201</td>
</tr>
<tr>
<td>14-1</td>
<td>Gathering Material for Models</td>
<td>208</td>
</tr>
<tr>
<td>14-2</td>
<td>Sources of Operational-Activity Information and Related Equivalency Models</td>
<td>209</td>
</tr>
</tbody>
</table>
14-3 Each Organization Level Is a Potential Source for Testable Descriptions as Well 211
14-4 Much of the Rhetoric is Written 215
14-5 Deriving the Testable Model 217
14-6 Different Types of Information that May Be Considered, Collected, or Created during EA 220

Tables

4-1 Steps in Obtaining a Measurement 61
5-1 Levels of Semantics in the Dog Program 82
5-2 Levels of Semantics in Mental Health and Schools 83
12-1 What Programs are Evaluable? 177
13-1 Examples of Universes, Domains, Flows, and Measurement 205
Several groups of analysts of government have long known a set of secrets that allows them to solve major problems of oversight and government operations when nearly everyone else is getting the wrong answers. The use of these secrets multiplies the reach of analysts of diverse background and helps pull together the teams of diverse specialists often needed to attack a particular problem. The secrets also provide a powerful set of questions against which to critique a study, evaluation, or policy analysis.

In this book we have tried to write down some problems that recur again and again in large organizations during day-to-day operations, including reorganization, reform, oversight, and evaluation. The approaches for surfacing these problems and dealing with them are the secrets. For a long time these secrets were viewed as quite dangerous by many people because they often quickly uncovered wide disparities between what everyone said was happening and what was really going on. As more people have turned to examining what the government can and cannot do and how the things that must be done are to be accomplished, the secrets have become more acceptable.

This book reveals the secrets of distinguishing more sharply between the necessary rhetorics of government and its realities. It tells how to compare the two, attempts to close the gap, and gives an introduction to using a special kind of model for representation. It also illustrates the importance of sequential purchase of information in dealing with large organizational systems. These and several associated secrets about how large organizations work are now revealed to you because, after all, you have bought the book.

We would like to acknowledge a large invisible college of colleagues, teachers, and friends in and out of government who—over the last ten years—tested ideas, contributed concepts, made applications of the material, and returned to tell us where our guidance had helped and where it had steered them wrong. A short history of the early development is given in the appendix, but to name all of the participants would be impossible.

We acknowledge more broadly all of those people in government who, when faced with nonsense or bewildering complexity, stop and—with or without this book—ask, “Is it possible to find out whether all of this activity is accomplishing its purpose—can it be evaluated?” They are natural evaluability assessors and are providing a wide base of acceptance for this work. We hope that this material will help them to continue to improve government.

The late Don Weidman (first with The Urban Institute’s Program Evaluation Group and then with the Office of Management and Budget) contributed not
only to the development of this work but also coined the term evaluable. The world is a little poorer without his work, wry wit, and accurate criticism.

Mary Sarley has been Joe Nay’s secretary as well as general overseer for this book for ten years, no easy task. More than that, however, she has been a colleague and friend who inexplicably refrained from braining both of us during successive revisions of this material over those years. She alone knows how many versions there have been.

The Urban Institute and the Ford Foundation funded some of the early developments of these ideas, even though some of the work was often threatening to cherished notions of how government should be run. Once, when the book was nearly abandoned, the Russell Sage Foundation called and asked us to apply for a grant to finish it. We never did apply for the grant, but thinking about it got us started again; for that, we thank Russell Sage. The National Rural Center provided some money for typing and revision of the draft preceding this one. The Performance Development Institute allowed this version to be typed as the final revisions were made.

The Experimental Technology Incentive Program (ETIP) Regulatory Processes and Effects Project at the Department of Commerce—a team effort over the last four years—has allowed us to correct many pieces of the theory that we thought we already understood. What we learned and accomplished there in regulation is the subject for some later book, but the book in your hand is much clearer and more accurate because of our experience in stretching oversight and assessment ideas from a program context to fit into regulatory-agency activities.

Joe Alan and Kenneth and Eric Nay proofread the last revisions. Warren Frederick of the Performance Development Institute provided invaluable conceptual help and examples of several concepts that could be fitted into this book. The format illustrations are due to his work. He is the only technical person (besides the authors) to have read this revision through from beginning to end. Any technical problems that have been missed should therefore be taken up directly with him. Robert Kershaw and his group at the General Accounting Office have made both theoretical and practical contributions to this work.

As the book took on its final form, our colleagues at the Performance Development Institute and at the Institute for Computer Sciences and Technology added invaluable refinements and insights, as did several evaluability assessors around town—in particular, Barry Rosenthal and Jim Statman of Aurora Associates and Marco Fiorello and Peter Eirich of Fiorello, Shaw Associates.

Finally, during the last revisions of chapters 9-15, we were provided with a sign over our basement work table by Eric Nay. It is from Soren Kierkegaard and reads: “All essential knowledge relates to existence, or only such knowledge as has an essential relationship to existence is essential knowledge.” Not bad advice for people searching for the actual locations of the direct interventions of government and for what is actually happening at those locations.
Eli Kay provided another prescient sign, to wit: “To err is human, but to really foul up, you need a computer.” Not bad advice for people searching for eternal truth.
This Page Intentionally Left Blank
Within hours of the announcement that the Reagan administration had appointed David Stockman to the post of director of the Office of Management and Budget (OMB), the gentle hum of copy machines and the clatter of collators was heard throughout the land, or at least throughout as much of the land as mattered to those affected first—that is, Washington, D.C. By the following morning, copies of a five-year-old Stockman article were on the desks of every senior bureaucrat and every important (and some not-so-important) consultants in the capitol area.

How this spontaneous generation of paper occurred remains a mystery to the authors, but occur it did and the news it brought was not good—namely, that this was no above-the-fray president talking; this was the man who had his hands on the money. Worse yet, he had been preparing a federal-spending hit list for at least five years. Hundreds of people moved to hundreds of phones and began to dial familiar numbers in the Congress—they paused as they recalled that their respective legislators had not returned to the Ninety-Seventh Congress. The iron triangles—tight coalitions of congressional committees, interest groups, and agency program directors drawn together to provide and maintain funding for particular programs—had been blasted asunder by the electorate.

Apart from its stunning effect on the bureaucracy and assorted feeders at the public trough, the Stockman article held some intrinsic interest. His major theme is illustrated by the following excerpt:

Having fed on a “starved public sector” rhetorical diet, the dominant liberal forces in Congress are simply unwilling or unable to recognize that this perhaps once appropriate complaint has ceased to reflect reality, and that a major reordering of...spending priorities...has now become imperative.

Stockman used several examples to illustrate the theme. The over-success of the Hill-Burton Act in closing the nation’s hospital gap was cited, as Stockman pointed out that the combination of a present excessive supply of beds combined with insurance-system incentives for in-hospital treatment had created a costly over-hospitalization of the U.S. public—that is, Hill-Burton had created excess hospital beds, and excess hospital beds had created their own excessive demand.

Stockman went on to say that the major programs of the great society, contrary to their creators’ intents, ultimately wound up subsidizing the haves rather than the have-nots. As examples, he mentioned several education initiatives and the once controversial community-action program that, he claimed, has been transmuted into innocuous social-spending pumps such as Head Start, Emergency Food and Medical Services, and Upward Bound.
This fact of transmutation of government programs brings us—albeit circuitously—to the point of this book. The dominant liberal forces in Congress certainly did not intend to subsidize the haves nor did they intend to bring the nation to the edge of bankruptcy in order to subsidize the have-nots. Whatever the intents, however, many observers believe that those results have nearly occurred. Whatever the intents of the newly dominant conservative forces, there is every reason to believe that—as Congress and the bureaucracy now operate—their intentions too will undergo similar transmutations.

Controlling the politically motivated transmutation of program intent depends, for the most part, on the self-discipline of the politicians in power (and those who do analysis for them), although procedural control mechanisms are available if the politicians choose to use them. For the most part, this book is not concerned with the deliberate transmutation of programs as they occur in Congress (although we take that up briefly in Part III and chapter 14). We do, however, investigate political motivations and program rhetoric, not from the aspects of rightness or wrongness but from the aspect of their relationship with program operation. Thus, setting aside for the moment the question of whether an emergency food and medical services program is an appropriate community-action-program enterprise, one of our major aims is to revive a long-dormant interest for any government operation in whether the intended intervention is being made at all, in the way that it was supposed to be carried out, and at a cost commensurate with its worth.

We found Stockman’s community-action-program example particularly intriguing because of one of the authors’ experiences with a small program agency in the South. A contractor team was investigating the operations of the local government agencies in a county of about 25,000 people. Of these, about 400 had been identified as being poor and living in substandard housing that was in need of weatherization. Conservatively figuring four people per substandard household, the number of housing units in question was, at maximum, 100. At the time the contractor team entered the county, the community-action agency was requesting additional funds from the three organizations (city, county, state) that had been jointly funding the weatherization program. The agency previously had hired a CETA-trained weatherizer and had secured sufficient money to finish the project if the weatherizer completed two houses per day—a projected completion rate based on experiences in nearby counties. Unfortunately, the weatherizer was completing only half a house a day; clearly, they needed funds for three more weatherizers.

When we explored the nature of the weatherizing operation, we discovered that the combined regulations of the three funding agencies compelled the weatherizer to (1) buy only enough material at one time to weatherize one house, (2) enter each purchase transaction on three separate and slightly different forms, and (3) stop in at headquarters to drop off each set of forms and pick up vouchers for the next batch of materials purchased. This to-and-fro traveling
ate up a large chunk of the weatherizer’s time (not to mention gasoline). Added to that was the time he spent filling out the required forms—a very long process since the man was barely literate and understandably made quite a lot of mistakes. In fact, the traveling and the clerking together took up almost exactly three-quarters of his time. All of which illustrates a fundamental principle of operation: It’s always more expensive when the carpenter types.

In regulation and defense, as well as in the social programs of this country, the same stultifying multiplicity of requirements is also occurring. It occurs because of different governmental requirements, but the effect can be the same on industry as on the carpenter. As regards regulation, for instance, the majority agrees that some regulation is necessary in many areas. Striking a proper balance of operation without having political intent transmute into undesired results remains as important and difficult as ever.

The evaluability-assessment procedures described in this book are based on the conviction that for government to operate sure-footedly, it must know what the problem being attacked is, how that problem can be handled, and how the suggested program operates in practice—not what someone thinks the problem is, claims will cure it, or how someone thinks the program is operating or intends it to operate. If the carpenters are spending most of their time typing, it is a sure bet that the houses will not get weatherized, and it is silly to spend money to discover that the people in them are still cold.

Enough experience has been gained to see that evaluability assessment will work with nearly any belief system or governmental operation. Indeed, the examples offered in the book range from social-service delivery to military helicopter sorties to garbage collection. In all cases, the principle is the same—namely, look at the intervention actually being made by the government. Vague rhetoric is usually a poor guide to policy formulation. The rhetoric of both conservatives and liberals often tends to differ from the reality of government operations in practice. Often, though, it is the rhetoric, rather than the reality, on which important belief structures are based. Money is spent, administrative systems are developed, and government operations are evaluated as if the rhetoric were the reality.

Distinguishing between belief structures and actuality and then making comparisons between them is essential to successful oversight and evaluation of government operations. In this book we present the evaluability-assessment method for doing so. Over the years, however, we have noticed that the federal government (like most large bureaucracies) has certain characteristics that may stymie even the most assiduous evaluability assessors as they attempt to draw the distinctions and make the comparisons. We therefore spend considerable time discussing those characteristics and offering some advice on coping with them.

This book is not tied to a given ideology. Liberals, conservatives, libertarians, and vegetarians, for that matter, can use the concepts and
procedures to good effect. People who believe they have an intervention that solves a thorny problem can use the techniques to demonstrate the rightness of their approach and to help carry out the solution (assuming that the solution is not demonstrably harebrained). People who believe they have discovered a program that is a ticking time bomb should, equally, be able to use the approach to demonstrate its dangers and to help dismantle it before it explodes (again assuming some plausibility to their beliefs).

We have tried to cover some of the problems that recurrently get in the way of conducting good oversight or evaluation. We begin, in Part 1, with some underlying ideas, address the representation of reality in Part II, and examine the people in charge of government operations in Part III. All of this is given as background for a presentation of the four cornerstones of evaluable assessment in Part IV:

1. The construction of two families of models—that is, testable models based on information derived from descriptions and equivalency models based on information derived from observation;
2. Comparisons and reconciliations within and between the two families to produce an evaluable model, often the basis for immediate action and always the basis of evaluation design;
3. The construction and use of functional models (see chapters 3 and 12) to display the relevant structure and flows of both the described and observed activities of interest;
4. A phased approach to the entire investigation that permits those in charge to make sequential purchases of information.

Earlier drafts of this book have been used—we hope to good effect—by a number of people. Drafts have been used as textbooks for students in masters of public administration programs; federal program managers have distributed them to contractors and bidders; government officials themselves have used them as primers; and congressional staff have read them in an effort to improve their oversight approaches. To all of those people, we apologize for any errors—either in concept or semantics—that appeared in the drafts. We think we have become smarter (or at least, more experienced) over the years.

Notes

3. This is, of course, a public-sector version of Say’s Law, the supplysider’s dictum: Supply creates its own demand.

5. The CETA program that trained him could hardly be faulted since it was supposed to produce rough carpenters, not office workers.
Part 1
Some Underlying Ideas

The material in this book is based on three underlying concepts:

Organizations are systems that exhibit purposeful behavior.

The existence and operation of information feedback loops and information comparisons motivate the purposeful behavior of organizations.

The stated rhetorical positions of management are often quite different from the activities that comprise the actual delivery of a specific governmental service or intervention.

Implicit in these concepts are the assumptions that evaluation activities are (or should be) (1) part of the organizational system, (2) an important (but not the only) information feedback loop, and (3) concerned with providing information about the actual delivery (as well as the rhetorical descriptions) of a governmental service or intervention.

Looked at in that way, it is clear that a book about doing evaluation cannot be a text about statistical techniques, a treatise on mathematical modeling, or a handbook on experimental and quasi-experimental designs. If it is to serve its function, a book about the evaluation of governmental programs should be about the operation of governmental organizations in general, with specific emphasis on how evaluation activities can most usefully relate to the rest of the system.

In order to understand the relationship of evaluation to the system in which it is embedded, the reader should have at least some familiarity with a few basic concepts about how organizations work. Therefore, the two chapters in this first part describe a simplified—or oversimplified—organizational context. Chapter 1 develops our definition of evaluation through an overview of the three domains within a government organization: (1) evaluation, (2) direct intervention, and (3) those in charge. Chapter 2 describes purposeful behavior and how it is affected by the operation of a single feedback loop. A simple, common system—a home heating system—is used to illustrate the points.

Any large organization is, of course, much more complicated than this. The root concepts developed in these two early chapters are expanded throughout the book to the stage that they can be used in actual organizations to solve real problems.
Evaluation in Perspective

Over the past several decades, evaluation has assumed an increasingly visible role in the operation of large government programs. Offices of planning and evaluation are part of the standing organizations of federal, state, and many local governmental units. Congress, in funding new programs, often designates an almost pro forma 1 percent for evaluation. Yet for all the attention paid to evaluation no universal consensus exists as to what it is and what it should do—whether it is a tool for deciding among alternative methods of delivering a given service or performing a given task, an offshoot of financial accountability, a way of rating programs long dead, or even the only way to spend 1 percent of a program budget.

Attempts at defining evaluation range from a few simple statements to entire books. Joseph Wholey et al., in *Federal Evaluation Policy*, provided one of the more-concise descriptions. Program evaluation, they said:

- Assesses the effectiveness of an ongoing program in achieving its objectives,
- Relies on the principles of research design to distinguish a program’s effect from those of other forces working in a situation,
- Aims at program improvement through a modification of current operations.

In other words, evaluation is a methodological approach to improve the quality of information about a program and to structure the information so that decision makers can use it while the program is still in operation. In this view, evaluation is part of purposeful management behavior. This book is based on this functional definition. The intended relationship that occurs when evaluation is part of purposeful management behavior is shown in figure 1-1.

Relating Evaluation to Purposeful Management Behavior

The bottom of figure 1-1 shows a process being carried out in its environment. A number of people are engaging in an activity or set of activities for the purpose of accomplishing a concrete objective—for example, to perform an appendectomy, to climb Mt. Everest, to win a ball game, or to distribute food stamps.
In order to direct purposefully the day-to-day operation of the process, the person immediately in charge requires some specific information, or measurements, based on the continuing performance of the process. For instance, if the process is a series of baseball games and the objective is to win them, the coach would want measured information including elements such as earned run averages, runs scored against the team, games won and lost, the team record in games played on natural versus artificial turf, and the record against various opponents. Gathering, analyzing, and reporting this information for the purpose of day-to-day program direction is one function of evaluation as shown by the internal loop on figure 1-1.

In addition to evaluation activity, nearly every real process is the focus of some additional activity that has been created by management. A baseball season, for instance, requires that players be assembled, uniforms bought, games scheduled, a ball park obtained, and so on. In most large operations, management is not directly involved in any of these real processes but, rather, oversees the activities from a distance. The more distant management is, the more often it bases its own activities and decisions, not on a contextual familiarity with a
process but on reports, gut feelings, and preconceived notions of what is going on out there. Together, this agglomeration of information and mental filtering form an abstract model of the process. In most cases, management has built and amended its model over time by trial and error.

The equivalence of management’s model to reality depends on many factors including the ability of the managers to accept reality, the reliability of reports, and of course, the size and complexity of the process, the complexity of the organization involved, and the number of people operating and managing the process. Even in small processes, management will always deal with some abstraction rather than with total reality since reality is so rich (and in part so irrelevant) as to disable decisions about it.

Another function of evaluation, then, is gathering, analyzing, and reporting information to management. This enables the managers to refine their models so that they are more nearly equivalent to reality and thus, presumably, a better basis for making management decisions. This process is shown by the outer loop in figure 1-1. Much of this book is devoted to describing this function of evaluation.

If figure 1-1 is divided into three domains (as in figure 1-2), the definition of evaluation can be more easily understood. One domain now contains the activities that are taking place in the process that is being managed. Another domain contains the activities of the people who are attempting to manage the process. The third domain contains the evaluation activity.

It is not hard to imagine, on an organizational level, cases for which some of the elements in the feedback loops are missing. A simple process could be operating nearly by itself. Certainly, many cases exist for which only the management and the process may be there. The managers may then control the process through their immediate sensing of what is occurring; they have absorbed the evaluation function into their own work. Or they may manage by only controlling inputs to the process and by paying little or no attention to how the process operates. This may be a perfectly satisfactory management method as long as the model the managers use has a sufficient resemblance to reality. However, if, for instance, the management model of how to win the world series is not sufficiently representative, decisions about training regimen, players to be traded, and so on may be in error. Successive measurements (for instance, game scores) may call this to management’s attention. To be successful, the managers will then alter their models of how their team wins ball games. Some form of evaluation activity (measurements and comparisons) is needed in order both to make further decisions and to compare expected results based on models of an activity with the observed actual results of the process itself.

It is not uncommon to find cases in which a process exists, evaluation
is ostensibly taking place, and yet no one is making any use of the results—sometimes because management is not organized, or motivated, in a way that allows the information to be used and sometimes because it is not evident that the information received is relevant to making decisions. This dangling evaluation occurs much more often in governmental organizations—partly because of unclear criteria—than in the private sector. In business, the profit-and-loss statement is an unavoidable evaluation measure. In sports, standing in the league is a clear, unambiguous measure. In government, however, readily available, unavoidable measures are much more difficult to obtain and agree upon. Further, there are less often directly involved vociferous aids such as irate fans or disgruntled stockholders with sufficient clout to ensure that evaluation results are examined and used to improve performance.

The primary purpose of this book is to examine some of the problems in developing agreed-upon measures, collecting relevant data, and ensuring the use of evaluation results by governmental management. The approach we use can also clarify those cases in which it is unlikely that the evaluator will be able to discover what is expected of the process or, even if expectations are articulated, where it is improbable that the intent can be realized. The evaluator needs, to be sensitive to the possibilities of both success and failure.
If Wholey et al.’s definition is now reviewed in terms of the simple organizational context shown in figure 1-2, “assesses the effectiveness of an on-going program in achieving its objectives” deals almost entirely with measurements and comparisons—that is, did the team play well? (Note that the definition of well involves the comparison of the measurements taken against a standard. Determination of that standard should take place in the management, not the process, sector. Did the team win the world series? If not, how close did it come?) “Relies on the principles of research design to distinguish a program’s effect from those of other forces working in a situation” is also principally a measurement-and-analysis problem with the emphasis on research methods to ensure that the evaluation is properly done—that is, did the team come into possession of the pennant because it won the series, or did the general manager win it arm wrestling the owner of the Kansas City Royals? “Aims at program improvement through a modification of current operations” is a statement about both comparison standards and the continuing use of evaluation information. A structure of potential information usable to management must be determined and designed and the evaluations carried out so that the desired information is produced and used.

The definition of evaluation can be reduced to three simple words: measurement (of what is actually going on), comparison (how the activity compares with some standard representing both the model and the expectations of management), and use (what gets done with the information). Analysis and comparison activities are carried out principally in the evaluation domain. Yet if management’s expectations are to be successfully measured, evaluators must go into the management domain in order to determine the expectations (uses for information are keyed to expectations) and into the process domain to develop models on which measurements can be based and, of course, to collect the basic data. The models and measurements should reflect the expectations of management. Suppose, for example, that management’s expectations are that the team shows up for the game looking presentable. In this case, management decisions would concern only the number and kind of uniforms to buy and various transportation arrangements. An analysis of won/lost statistics would be irrelevant. The attempt to design the proper measures and comparisons and to obtain usage requires much information from, and possibly some preconditions being met in, both the process and management domains.

A detailed discussion of the domains is beyond the scope of this introductory chapter. However, a quick overview is in order.

**Exploring Organizational Boundaries**

In order to describe evaluation in terms of its effect on organizational
behavior (and the effects of the organization on evaluation), it is necessary to develop some abstractions of our own that have close parallels in reality. The organizational reality is often as impenetrable as a thorn thicket. We can, however, develop a simplified framework sufficiently realistic to permit an examination of the organizational and evaluation processes.

We have indicated that at least three domains are important: evaluation, direct intervention, and those in charge. (Hereafter, the rather awkward appellation those in charge will be used instead of management since many of the activities in the government do not in any way resemble what is commonly called management by professional managers.) Those three activities are arranged as shown in figure 1-3 and affect each other (in terms of short-term operation) as shown by the arrows. The activities considered to be in each area of the diagram are described in later sections.

The reason for exploring each of these domains is simple enough. In actually laying out a useful and usable evaluation, information from each of the sectors must be gathered and brought together. Most of the information to be gathered is particular to, or may be thought to be from, one of the three domains of interest as developed here. The people who inhabit the different domains have different perspectives and different needs. They even speak slightly different languages. For the most part, the people in the evaluation and those-in-charge domains are dealing with more-abstract models of the real process in question. The direct-intervention domain actually contains the real process. It is a matter of some importance to be able to identify the domain from which different types of information are gathered, since without that demarcation it is difficult to distinguish the degree to which such information represents actual events.

**Evaluation**

We assume that true evaluators must do one or more of the following things:

1. Construct models for use in measurement and analysis. These models are simply diagrams of some sort that display how the characteristics of the process to be measured are assumed to be interrelated and the importance of the
characteristics to each other, to the points of measurement, and to the environment. They reflect an informed abstraction of reality that is tailored to the uses of those in charge.

2. Make measurements of some part of the intervention proper or of some activity or phenomenon explicitly assumed to cause (or enable) or to be caused by the intervention. For instance, runs scored is a measurement taken from the intervention, uniforms ordered is a measurement of an enabling activity, and fan satisfaction is the measurement of a phenomenon in the environment assumed to be caused by the intervention. Assumptions about causal relations should always be explicitly and carefully stated—that is, the assumption that, if the team plays well, its performance will cause the fans to be satisfied, is different from assuming that satisfied fans are an indication that the team is playing well. If you doubt this, remember that the newly formed New York Mets played to packed houses of ecstatic fans when the team was unquestionably the worst in baseball. By the same token, the world-champion Oakland team performed in a near empty stadium.

3. Perform data analysis to bound the reliability of the measurements taken. For instance, if runs scored and runs scored against are selected as measures of the quality of team play, how many games must be analyzed in order to be 70 percent sure of the answer?

4. Analyze sets of related measurements in order to test the validity of the model being used to represent reality—that is, are the displayed characteristics really as interrelated and important as believed? For instance, suppose that runs scored and runs scored against are measured in every game of the season and that runs scored outnumber runs scored against by two to one. Because all the games have been analyzed, it can be inferred that the measurement is 100 percent accurate. If it then turns out that the team has lost four-fifths of its games, it is a reasonable inference that either a very rare event has occurred or that something is wrong with the measurement model—that the assumed interrelationship between runs scored/runs scored against and winning is invalid.

5. Compare the models of the real process, on which measurements are based, with the models constructed by those in charge, on which expectations are based. These comparisons might cover a range of questions such as: Is there a group of appropriately equipped men playing ball? Is the new defense combination working? Is the activity directed toward the objectives of those in charge of the operation? (for example, is the team’s won/lost record good enough to reach the world series?)

6. Reduce the results of any of these preceding steps to two forms: One form, easily readable by a technically literate reader; the other, by busy senior people.

Many other activities may be done by evaluators. For example, they may write guidelines; give talks on methods, goals, and objectives; draw
organization charts; or spend endless hours in discussions with deputy assistant secretaries. Unless they are also doing at least one of the six items just listed, however, they are not doing what we define as evaluations. As described here, evaluators get their information from two places. Concepts, grand plans, and goals (which contain management’s expectations) are obtained in interviews with those in charge. Measurement data of the process and its outcomes are more usually taken in and around the direct-intervention domain. This domain is described next.

**Direct Intervention**

The purpose of many interventions by government is to deliver or perform a service of some kind or to alter the way in which a service is performed by others. A direct intervention, as used here, is the actual delivery or performance of the service. It does not include the policy decisions about the service or any of the other myriad activities that are predicated on someone’s abstraction of the process.

The point of intervention is defined as the boundary between the person delivering the service and the recipient of the service. For instance, the government employee who actually does something for or to a citizen is at the point of intervention (for example, an employment-service counselor who places a citizen in a job, an army sergeant who trains a recruit, a police officer who arrests a burglar, an employee of the sanitation department who picks up the garbage, or an IRS employee who reviews and examines income-tax returns). In some cases, the personnel at the point of intervention are not government employees per se but people who have been commissioned (usually by contract or grant) to perform the service. Thus, if a city chooses to hire a garbage-collection firm rather than have city employees pick up the debris, the garbage collectors are nonetheless the people at the points of intervention. The people who hired them to do the dirty work are not in the direct-intervention domain. This is an elementary distinction but an important one. To restate, the point of direct intervention is the location at which the performance of service actually takes place. (A model of the process of direct intervention constructed for analytic use is an abstraction. Measurements can only be made of the characteristics of the reality, not of the model. If an evaluator is to measure an intervention activity, the evaluator must go to where the action is—namely, the point of intervention.)

Figure 1-4 is an expanded representation of a direct intervention embedded in its environment. It shows the different places at which measurements of a direct intervention can be made.

The model shows the direct intervention sitting in its immediate environment. The direct-intervention domain is peopled by everyone directly effecting or directly affected by the intervention. Normally, these include the
immediate supervisor of the intervenor as well—for example, a baseball coach. This domain is composed of people who deal principally with the process under examination. Those who work mainly from a model of the process are not included in the direct-intervention domain.

The cloudlike outline in figure 1-4 represents the boundary of the intervention’s immediate environment. The inputs to the intervention are drawn from within the boundary. The inputs to the intervention are those things that will be directly affected by the intervention (for example, garbage, people, potholes).

Contributions to the intervention process are things that are intended to (and sometimes actually do) help the intervention take place. These inputs include money, uniforms, guidelines, technical assistance, and intervenors.

Process measures describe how the intervention is being carried out and to what extent. These measures concern how the operation is getting on without regard to the overall effects it may have. Process measures often include things such as the action taken by the people involved, how many people are serviced, what exactly a service consists of, how convenient the arrangements are, and how people feel about them.

Outcome measures are, in effect, the last easily attributable process measures. They describe how given characteristics of the input were directly altered as an end result of the intervention in a way directly attributable to the intervention.
Impact measures describe the effects of the intervention and its outcomes on the environment. Impact usually involves the test and validation of some cause-and-effect hypothesis. An accurate impact measure is often a contradiction in terms since it implies that attribution or demonstration of cause and effect can be established—usually a dubious demonstration since the expected impact almost invariably takes place at a temporal and/or logical distance from the intervention.

The difference between process, outcome, and impact measures is sometimes ambiguous near the edges. For instance, time spent by professionals is clearly a process measure. Number of people trained to be welders as a measure of a job training program could be either a process or an outcome measure depending on the goal of the project—for example, is it supposed to train welders, to place people in jobs, to raise people’s incomes? Thus, the categorization of particular measures as process or outcome often represents an interpretation of expectations. The interpretation is usually derived through a process of iterative interviews with the intervenors and those in charge.

Earlier, we stated the assumption that if the team plays well the fans will be satisfied. The play of the team during a baseball season could be regarded as an intervention; the players then are one of the inputs to the intervention; bases are contributions to the process; runs scored is a process measure; and the team’s standing in the league is an outcome measure. So far, the measurements have been reasonably precise and the interrelationships among them and the activity reasonably straightforward. The fans, however, are out in the environment. Their relationship to the intervention is shrouded by unknowns. One may feel that if the team plays well the fans should be satisfied. That moral imperative may then lead one to assume that if the team plays well the fans will be satisfied. The jump from the moral imperative to the rational certitude is, to some extent, a blind leap, and a number of intervening variables may appear while the evaluator is in midair. Fans may be dissatisfied because the concession stand sells warm beer. Fans may be satisfied because the manager often kicks dirt on the umpires. And what do we mean by fans? Are they people who pay for game tickets? Who watch the team on television? Or are they all the people in town that might be enticed to watch the games if the team played well enough? Despite the tenuous relationships between the intervention and its impact on the environment, evaluators are often expected to measure the impact and often must model and include the environment presumed to be affected, relating the various measurements to each other. When selecting impact characteristics to be measured, particular care must be taken to choose characteristics that
fanatical care must be taken to lay out the entire chain of assumptions about cause and effect, the location and nature of possible intervening variables, and the measurements that serve as adequate proof of the assumption chain.

Control or comparison measures are sometimes made. One purpose of these measures is to control for intervening variables—that is, to gain more assurance that the observed change in the inputs to the intervention, or in the environment, is really caused by the intervention. The control or comparison measurements are taken from some group or process that is believed to be similar to the one being evaluated but that has not been affected by the intervention. The two sets of measurements are compared, and if a change occurs in the group or process being evaluated but does not occur in the control or comparison group, the change is then often attributed to the intervention. The selection of control or comparison groups is a complicated business and historically seems to involve nearly as many assumptions as the selection of impact measures. The history of the use of comparison groups in social-sciences research contains many more questionable cases and failures than unambiguous successes. As in the selection and reporting of impact measures, the evaluator should proceed warily and explicitly.

**Those in Charge**

As a rule, direct interventions by government do not just happen. Somewhere, sometime there were enabling interventions, or interventions intending to create direct interventions. These enabling interventions emanate from some source of authority—for example, Congress, a city council, or a school board. Sometimes the enabling intervention takes the form of an explicit directive—for example, “The public works department shall build a bus terminal at M St. and 21st.” Sometimes, the enabling legislation is little more than an expression of good intent—for example, “The ombudsman shall ensure that all citizens dealing with the city government be treated fairly.” Quite often, the enabling intervention is an amalgam of political compromises and encompasses an astounding number of hopes and dreams. The those-in-charge domain lies between the source of authority and the direct-intervention domain. Those in charge are the people who translate the language and intent of the enabling intervention into directions and guidelines for the direct intervenors and who pass along the money.

This sector is especially interesting in large bureaucratic organizations. In many cases, the actual language of the enabling intervention was produced either as the result of a political compromise or of an intent to go forth and do good. In these cases, do not be surprised to discover that
quite extended chains of effects are assumed (at least rhetorically) to be caused by even the simplest of enabling interventions.

For instance, funding a day-care center is supposed to cause a series of outcomes and impacts including, but not limited to, the following:

- Replacing inadequate child care with adequate child care,
- Enabling children to live up to their potentials,
- Raising the net income of the poor,
- Reducing welfare roles
- Disencumbering women so that they can enter the labor force.

One way or another, those in charge must translate these expectations into directives that presumably guide a real day-care center.

The those-in-charge domain includes more people than those in a straight line between the authority and the direct intervention. It also includes people who are owned by those in charge. For instance, some people act as in-house extensions of those in charge—for example, secretaries, assistants, office managers, or vice-presidents for acquisitions. It also includes people who operate the ancillary activities to the direct intervention. A baseball team, for instance, requires people who purchase uniforms, rent ball parks, or sell tickets.

Therefore, those in charge are usually operating on two levels. On one level they are involved in a real process (albeit not the process that principally concerns the evaluator). They write letters, answer the telephone, mediate (or cause) office disputes, and hire and fire subordinates.

On the other level, those in charge deal with their models of the real process (or direct intervention) under consideration. These are called testable, or rhetorical, models for the sake of distinguishing between them and the equivalency models constructed for measurement and analysis purposes by the evaluators. On the basis of the rhetorical models, those in charge provide certain contributions to the process—for example, money, guidelines, and supplies. It is not uncommon to find that, among those in charge, several different rhetorical models are in use for the same direct intervention. Figure 1—5 helps to illustrate why this happens.

In a large bureaucratic organization, usually several layers of management exist between the source of authority and the direct intervention. Each layer deals with its own activities, has concourse with different units of ancillary activity, and ordinarily has its own perspective. For instance, the person who administers the financial aspects of the contract for a day-care center has a much different testable model than the person who is responsible for enforcement of federal day-care standards. Both models will be considerably different from that of the executive who operates on the policy level. It is possible that none of the three models will look much like the day-care center that actually has children in it.
There may be a few organizations in the world wherein the management team plans, organizes, coordinates, and controls. Based on our experience, such organizations are not likely to contain those in charge of a complex government program. The managers usually are too busy for such activities. Their days are spent cajoling subcommittees or vendors, sitting on advisory committees or being advised, talking with people outside the organization who may have important political information, giving information to media representatives, coping with the endless administrivia visited upon them, or just reacting to crises. Very little then is left to reconcile the disparate testable models that exist either in the political domain or the those-in-charge sector or even to test their expectations in great detail. Any systematic planning that occurs is often done off the record by a long-time confidant. Occasionally such work may be attempted in an office entitled planning and evaluation—usually one of the more-arcane units of government.

![Figure 1-5. Those in Charge](image)
What, then, Is an Evaluator to Do?

In order carry out the three-part process of program evaluation described at the beginning of this chapter, the evaluator must extract the expectations about the intervention from those in charge. The testable models must be examined, reconciled, and reduced to testable terms. By testable, we mean a coherent description of an assumed process that can be compared to the process as it actually exists. Choosing the precise testable models to work with is not a simple matter. Virtually all the members of the those-in-charge domain have their own models of both what they do and what others, including the direct intervenors, do. For the evaluator, an important and often unavoidable testable model is the one owned by the person who is requesting the evaluation, especially if that person is likely to use the results. It is not, however, the only important model. To select the important models it is necessary to discover who will be involved in implementing the results of the evaluation as well as what the intervention process actually looks like. The official chain of authority is another good starting place because tracing it out often uncovers the actual lines of authority. Another good plan is to follow the money—that is, if a change is to be made, the hands that make the change often hold the dollars.

For example, the testable model of the person actually owning the ball club may be a very important one. The person in the ball club who is responsible for handing money over to the general manager also has a model worth noting. The general manager, in turn, allocates a certain amount of money to the coach. The coach’s model is clearly of some significance. (Whether the coach, or the first-line supervisor, belongs in the those-in-charge domain, the direct-intervention domain, or somewhere in between depends on the way that the coach performs the job. Some so-called office coaches work entirely from models of the activity; others deal directly with the reality; and most do a bit of both—for example, work with the offense and leave the defense to an assistant working with some general guidance.)

Recognize, though, that the routes of money and authority are only two of several possibilities. In the event, it may prove more fruitful to trace the models of the people who transfer things such as information or influence. The only way to tell is to choose a promising path, follow it, and see where it leads.

Some of the ancillary models may also be important. If an assumption implicit in the predominant testable model is that good equipment is essential to good play, it would be of no little interest that the person responsible for buying equipment works from a model in which spiked shoes are purchased in bulk from a cut-rate outlet to hold down expenses. It is imperative that the evaluator talk to enough people among those in, charge to begin to construct a unified testable model that shows what the program is believed to look like
and what the specific expectations for it are.

It should be noted that the actual expectations for a program are often buried deep in the rhetoric. Mistaking the nature of expectations can lead to fundamental mistakes in the construction of the measurements to be taken. For instance, suppose that the expectation for the ball team was stated as “Win the world series.” Rhetoric aside, the objective might really be to make money. If that turns out to be the case, the boundary of the direct-intervention sector must be redrawn to include all of the money-making activities such as concession rental; some personnel previously thought of as ancillary to those in charge now become intervenors (for example, ticket sellers); and winning the world series is now regarded as a process rather than an outcome measure. Uncovering the expectations can be a laborious process involving frequent conversations with both those in charge and the direct intervenors. The evaluators should be prepared to scrap their models and measurements and to start over if it becomes apparent that the originals were based on mistaken understandings of the expectations. Evaluators are not the only people who misunderstand the expectation of those in charge. Successful coaches sometimes lose their jobs that way.

The rhetorical information collected from those in charge should be used in preparing a unified testable model of expectations—that is, one that describes functional relationships with measurable expectations of outcome. This testable model is a model of the beliefs of those in charge that can be reality tested through comparison with a model of the process based on the direct observations of the evaluator.

The first test check is to see if this testable model is fiction or fact. The evaluator does this by personally going and looking at the actual process in question. Quite often surprises await. The testable model may be one that depicts an activity designed to train people for gainful employment with the expectation of placing them in long-term jobs where they perform services desperately needed by grateful communities. The reality may be that thirty currently unemployed hairdressers are trained each month in a community glutted with beauty parlors. This initial information may be all the busy people in charge want to know, or it may be something hardly anyone wants to know at all. However, bringing expectations and measurements together always provokes interest. Either reaction or ignorance may result. (The act of ignoring hard information that contradicts expectations and rhetorical positions is, of course, one of the two definitions of management or legislative oversight.) The evaluators have done the first step of their job. Those in charge can now grapple with whether to change what they believe about the process or to change the process itself. If more information is needed—if, for instance, the process actually looks something like the testable model and if those in charge want to know how close the intervention is coming to their expectations—then the evaluators attempt to
construct more-detailed models that are equivalent to the direct-intervention sector and to select measurements based on these models. Matching the expectations (in the form of testable, or answerable, questions) to measurable phenomena (at the direct intervention), selecting the measurement instruments and comparison methods, and refining the evaluation design are the next steps. Evaluators can then go forth and make measurements, analyze their data, validate their models, assess the meeting of expectations, and prepare their reports with some hope that the effort will lead to program improvement through a modification of current operations. In fact, the preparatory steps described here often lead to program modifications before further evaluation work is done.

Note

Chapter 1 placed evaluation and the organization in perspective. Evaluation is only one aid to an organization in achieving purposeful management behavior. This chapter expands on the nature of purposeful behavior itself and evaluation as a part of that behavior.

Behavior is purposeful when it is directed toward a goal and when the attempts to close the gap between actual performance and expectations for that performance are predicated on the size, nature, and tendencies of past and present errors in meeting expectations. The expectation can be conscious, in that a policy decision has been made to attempt to do something, or unconscious, like the habit of placing one foot in front of the other in order to walk. Whether conscious or unconscious, an expectation exists, an attempt is made to attain it, the error between expectation and performance is sensed, and behavior is adjusted so as to reduce the error.

Herbert Simon has pointed out that “the simplest movement—taking a step, focusing the eyes on an object—is purposive in nature, and only gradually develops in the infant from its earliest random movements. In achieving the integration the human being...observes the consequences of his movements and adjusts them to achieve the desired purpose.” Infants learning to walk do several things. They decide what they want to do (initially, this is only a policy). They then attempt to do it. As they try, they compare the results of what they just did with their standard of proper performance. Later, a little older and wiser, they try again, each time modifying their behavior in order to reduce the error between what happens and what is desired.

Thus, the child, learning to feed itself with a spoon, may initially put food all over itself (and its environment). In successive iterations, however, the child learns to focus control of the process directly upon the error distance between the spoon and the mouth and continues to improve the ability to reduce this error in actual practice until eating with a spoon becomes a learned behavior stored away in the brain and usable whenever needed.

Purposeful management behavior has many similarities. Management expectations are often initially policy decisions. Through organization, managers attempt to turn their expectations into reality by bringing to bear the parts of the organization that should be able to carry out the policy. The hard part comes in getting the workers to compare the results of what they have done with management’s expectation for successful accomplishments and, if necessary, to
alter their (the workers’ own) behavior so that, over time, the gap between expectation and performance continually narrows.

The four essential elements of purposeful behavior are

1. **Expectations**: Standards with which the desired performance can be compared. These may be absolute standards, standards relative to performance of others, or even present performance (do better or no worse than last year);

2. **Performance measurement**: A measurement system that reliably determines the results of the actual efforts in the same terms as the standards;

3. **Comparison**: Comparison of measurements of performance to the standards (or expectations) in order to determine the actual errors in meeting those expectations;

4. **Responsiveness to comparison results**: A part of the organization that is able to create the activity desired and willing to alter (often continuously) its approach, effort, and technique over time. The alterations, based upon the errors found in the comparison, should continue until the distance between actual and expected performance becomes acceptably small.

Evaluation provides the measurements and comparisons for such feedback and is often also involved in translating the expectations into clear standards. Used in support of purposeful management behavior, the functions of evaluation are to create methodologically sound information in a manner that permits valid comparisons with a standard, to perform those comparisons, and to inform the managers and the operators of the results of the comparisons.

A large (and still growing) body of literature exists describing sound methodologies that involve sophisticated (or unsophisticated) statistical techniques and experimental designs, usually dealing with the problem of obtaining useful measures and comparisons out of already sound data. However, if the implementation of those designs and the exercise of those techniques are to support purposeful management behavior (rather than to exist as random activities or academic exercises), the evaluators and managers must understand the several steps in the creation and operation of successful performance feedback systems.

Members of any organization participate in many groups both inside and outside their organization. As Chester Barnard has pointed out, this participation conditions the actions they take. Members of an organization receive feedback from many places. Their actions are compared to the expectations of friends, supervisors, and colleagues and
expanded upon Evaluation

returned to them as praise, arguments, or attacks. Even the simplest of social organizations is a literal maze of performance feedback loops. In a smooth, well-functioning organization, the majority of these control systems guides the organization in a single direction toward a compatible set of standards. Other organizations display a high degree of schizophrenia as their members respond to discordant error signals. This schizophrenia is not uncommon in large government organizations that receive signals from the Congress, the current administration, constituencies developed under past administrations, bureaucratic superiors, and so on.

If the evaluators are to produce results that are used, rather than filed, it is essential that they recognize and understand the important feedback systems operating simultaneously in the organizations being evaluated. However, before attempting to cope with these multiple loops, it is necessary to understand how a single loop operates.

Therefore, the remainder of this chapter describes a single performance feedback system as it operates in the simple, mechanistic milieu of a home heating system. This illustration is used because it is both relatively straightforward and familiar to most people. Note though, that this example is not an accurate model of a complex social system. It is offered only as a useful first step toward understanding the more-complicated phenomena.

**Operation of a Simple, Mechanistic Feedback System**

A home heating system is an illustration of a feedback system managed by information based upon measurement and comparison. Figure 2-1 is a schematic (following the paradigm of purposeful behavior presented in chapter 1) of the operation of a home heating system. The figure displays some of the essential operations that take place.

The those-in-charge domain occupies the upper left-hand corner of the figure. Even in as simple an operation as a home heating system, this domain is conceptually complex, and we expand on it in detail later in the chapter. Suffice it to say here that, in a home heating system, the control mechanism that issues off/on instructions for the furnace is located in the domain of those in charge.

The lower left-hand corner of figure 2-1 represents the direct-intervention domain. It contains an oil tank, a furnace for burning the oil, a circulating system for moving hot water to a radiator, and a radiator that radiates the heat. Also in the direct-intervention domain—but not of it—is a temperature sensor that measures room temperature and reports the data to the evaluation domain.

The operation of the direct-intervention domain normally follows a well-established routine governed by predetermined rules. On receiving a signal from the administration, the furnace turns on and burns the oil,
Figure 2-1. A Home Heating System
turning it into hot gas; the hot gas heats the water; the heated water circulates through the radiator; the radiator radiates the heat into the room; the house gets warm. On receiving another signal from the administration, the furnace turns off. The purpose of this activity is to keep the house at a comfortable temperature. To this end, the furnace receives its stop/go directives from those in charge and proceeds with its own established implementation routines.

An important thing to note about the direct-intervention domain is that, in this case, it has no means of evaluating its own performance. In the absence of the temperature sensor, comparisons, and the feedback of the resulting information (if, say, the automatic administration had been programmed to turn the furnace on and off in half-hour cycles), the furnace would simply go on mindlessly repeating the sequence of operation that it knows best each time it received a go signal. It would do so if there were icicles in the living room, and it would do so if the house were on fire.

The additional elements that enable purposeful behavior to take place are in the evaluation domain, which is illustrated in the remaining portion of figure 2-1.

The evaluation domain is linked to the those-in-charge domain through information about whether expectations are being met. The temperature sensor (an evaluator) measures the temperature of the living room, and the thermostat (another evaluator) makes comparisons between actual and expected living-room temperature. The those-in-charge domain receives the results. The sensor measures the most relevant output of the furnace to the inhabitant—namely, the temperature of the room. The temperature of the room is evaluated by the thermostat by comparing the present room temperature to the expected temperature set by a higher level of policy management.

Because most house-temperature management is not really concerned with exact temperature but only with a range of acceptable variations, the evaluation comparison is not reported to the administration unless the heat level observed by the sensor exceeds the limits of permissible error. When the room temperature hits the upper boundary of the permissible range, the evaluation comparison signal is transmitted to the administration and the furnace is turned off. When the room temperature hits the lower boundary of the range, the furnace is turned on. The room temperature measured by the thermostat is also displayed (reported) to those in charge on an agreed-upon scale divided into well-defined, well-known units (for example, degrees centigrade or Fahrenheit). This display of numerical temperature measurement (shown on the thermostat) is not necessary for proper operation, however, since administrative control of the furnace operates from the error signal generated by comparing the temperature measured by the sensor with the expectation set on the
thermostat. This comparison would still be made and signaling would still occur even if the printed degree scale came loose and fell off. The degree scale is there simply for the record—to provide an indication that the room temperature has been requested to be somewhere in the vicinity of the temperature shown.

All of the elements of purposeful behavior are present in the case of the furnace:

1. **Expectation** A small standard range around an acceptable room temperature established by setting the thermostat,

2. **Performance measurement:** Measurement of the room temperature in the same terms;

3. **Comparison:** A comparison of expected and measured temperature that generates an error signal (including direction of the errors) when expectations are not met;

4. **Responsiveness to comparison results:** The control circuitry and, of course, the furnace itself that respond to the error information in order to keep room temperature within expectation over a wide range of conditions.

What is described here is one very simple arrangement that accomplishes purposeful behavior by the use of these four elements. The thermostat in a home does not have to be very accurate because its actual setting can be controlled by an additional feedback loop. Figure 2-2 shows a policy function added in the those-in-charge domain. The policymakers determine the proper temperature in the room and announce it by setting (or resetting) the comparison standard (the thermostat).

As in figure 2-1, policy directives concerning room temperature are implemented by the heating system. The furnace is located in the direct-intervention domain. An evaluation system and an automatically implemented set of administrative procedures in the those-in-charge domain are also present. This purposeful-behavior loop has a clear, if implicit, charge: The furnace is to be used to keep the house at a comfortable temperature.

Those in charge engage in certain activities. Some in-house activities may have nothing to do with the actual furnace operation—for example, painting the furnace blue or building a redwood box to hold the thermostat. There will certainly be some ancillary activities—for example, ordering fuel and obtaining storm windows. Also, some oversight of the actual process of heating a home will occur. For the most part, however, heating of the home will still take place through the automatic administration or predetermined rules—that is, once the furnace has been installed and a comparison standard decided (a major
Expanding upon Evaluation
policy decision communicated to the purposeful system through the device of setting the thermostat), the operation proceeds automatically so long as nothing unexpected happens. The furnace goes on and off and heats the room according to predetermined rules in response to directions based on information about the temperature in the house. Only when the manager discovers that something out of the ordinary is happening do managers—as distinct from administrators—do something.

If the house gets too cold or too warm for comfort, management checks the setting on the thermostat. If the thermostat is indeed set to reflect the predetermined wishes of management, then the actual temperature of the room is checked. If the temperature is within the normal range of error, management then decides that the predetermined standard is wrong (a policy decision) and (by raising or lowering the setting as desired) directs the system to use a new comparison standard. The administration will then act on error signals from the comparisons made against the new standard and raise or lower the temperature of the room accordingly.

Thus, if management should decide that the predetermined standard is wrong—if the house is usually either too hot or too cold—management will reset the comparison position of the thermostat, shifting the entire band of comparison up or down.

If, on checking, it is determined that the standards do represent a comfortable temperature level but that the furnace is not approaching the standard—that is, the temperature is not within the normal range of error—then management does a quick check to make sure that nothing has interfered with the evaluation/administration/process loop, such as no oil in the furnace or a cold draft blowing through an open window onto the sensor. If nothing easily correctable is apparent, management then issues a direct order to the administration to alter the furnace’s behavior forthwith. The order might be transmitted via the on/off switch. The furnace would then be closely managed (that is, management manually operates the on/off switch) until repairs are made and automatic administration can resume. During the period of close management, the managers, by doing their own measurement, comparisons, and actions, have replaced (or become) the loop that controls the furnace. An important point to note is that if management is to effectively control the furnace by using the on/off switch, the thermostat must be disconnected (that is, automatic administration must be stopped). Picture what could happen if the automatic system was keeping the house too cold. After verifying the thermostat setting, the management response might be to go to direct control. In that event, management would flip the switch to on, activating the furnace. However, the furnace would still be turned off when the room temperature reached the limit acceptable to the sensor. This may happen long before management is satisfied with the
temperature.

This superficial description of the management of a home heating system is sufficient to illustrate an important concept. The those-in-charge, direct-intervention, and evaluation domains are linked together to form a stable feedback system. An overlapping control loop—a policy-control loop that changes the comparison standard—can be activated when those in charge want to alter the overall outcome of the process. The alteration in expected outcome can be effectively accomplished, however, only if management recognizes, understands, and manipulates the basic operating loop on its own terms. This can be done by:

Taking advantage of the stable loop and altering performance through policy—for example, changing the comparison standard that governs the furnace’s activities and letting the existing measurement, comparison, and response arrangement produce the new conditions;

Entering the existing loop and directly controlling the actions of the administration—for example, disconnecting the thermostat and manually operating the on/off switch.

Management may also have to deal with problems beyond the control or competence of the purposeful administrative system by:

Meeting necessary operating conditions—for example, putting oil in the tank;

Removing noise from the feedback loop or system—for example, closing the windows.

It is unlikely that behavior will be altered, as desired, if the existence and operation of the operating feedback loop is ignored. For instance, painting the furnace white instead of blue will not make the room get cooler, nor will manually operating the on/off switch be a satisfactory long-term solution. One way or another, management must deal with even a single automatic-feedback loop on its own terms or else redesign it.

Obviously, a home heating system could be operated in a number of different ways. It is conceivable, for instance, that management would demand a virtually absolute standard with an imperceptible error range. For instance, management may be trying to comply with a fuel-saving policy of maintaining a temperature of exactly 65°. It is possible to meet an absolute standard. However, the pleasure derived from such pinpoint measurement is not normally deemed—sufficient to compensate for the pain in the pocketbook. Usually, the more accurate the control instrument, the more expensive it is. That is, of course, also true in organizations.

Alternatively, the system could be operated as an open loop (no feedback)— for example, by allowing it to burn only so much oil every
hour or by turning it on and off in predetermined time cycles. This would eliminate the feature of controlling from the sensed error between room temperature and a standard. This closely resembles running a government agency by attempting to control its allocations and expenditures. More managerial oversight also could be used (another common solution), and management could sit by and control the furnace directly at all times.

The actual system used in the home has evolved over time, however, and has been found to be essentially satisfactory, reliable, economical, and unobtrusive.

**Why Is Life Like a Furnace?**

In many respects, life is not like a furnace. Such a mechanical explanation of the behavior of organizations and the people in them is not only repugnant but also oversimplified to the point of absurdity. Even when a governmental manager has the wit to check for noise in one of the program’s feedback loops, it is likely that the noise is generated by yet another loop (like an angry congressional appropriations committee, a school board or even by citizens). Administrators do not always run through their routines regardless of ice in the kitchen or fire in the basement (it only seems that way). Direct intervenors often can sense for themselves when something is wrong with the operation. What is more, they sometimes tell those in charge about it.

Despite the differences, however, certain fundamental similarities exist between a furnace and each of the many feedback loops in a complex social organization. Primarily, feedback-loop behavior is pervasive, and the tendency to control and operate on some kinds of error signals is common. While the rhetoric of social organizations often tends to be obfuscatory—sometimes deliberately so—many basic loops and comparisons are almost always there.

Most midlevel bureaucrats respond to many feedback loops whose signals, comparison standards, and even directions may change. An important step is to determine what the existing formal and informal organizational control systems are whenever a new evaluation-design problem is to be approached. Somewhere, someone is doing something. Somewhere up above someone else can tinker with the system or with the standards of comparison. Somehow information created from measurements and comparisons (formal or informal, true or false, useful or irrelevant) gets back to the people who can tinker. If the evaluator can identify the loops and their elements at the beginning of an attempted new design, then the evaluator has at the very least identified many of the people who have important questions that will need to be answered, beliefs that will need to be examined, and process information that will need to be obtained. It should then be possible, for instance, to avoid wasting time and money to determine the program-evaluation equivalent
of what color to paint the furnace.

Established information loops and behavior will often exhibit an amazing stability in the face of additional information unless that information can be inserted through an existing accepted feedback loop or accepted new loops can be created. In constructing a new feedback system, remember that four conditions are necessary for any working loop; accepted expectation standards, measurement, comparison of expectation with performance, and a willingness and ability to act on the resulting information.

The identification, use, and (when necessary), creation of feedback loops is usually regarded as beyond the purview of the evaluation team. However, our observation has been that if the evaluation team does not include these organization-analysis tasks in its work, then no one will do it. The evaluator should either plan to be involved in such work or to expect much of the evaluation work to be wasted.\(^3\)

Many examples of simple feedback loops could have been chosen to illustrate the operation. The furnace was used because most people intuitively know how it works and because it clearly illustrates several properties of research design: the selection of a characteristic to measure, measures and a measurement instrument, the point at which measurement is to be made, the comparison to be made, the error from the comparison as a signal for action, and the range of desired accuracy. It is interesting to note how the presence of real-time feedback about a desired or expected value of a characteristic simplifies the gathering of data, the production of information, and ultimately, attaining expectations. It is also interesting to note that the real-time feedback system is concerned with the most relevant expectation (comfort) and that the administrative routines are based on a thorough knowledge of how the system works. For instance, when the house is too hot, the furnace is turned off—the oil is not drained from the tank.

It is important to understand the actual intervention being made and to have models equivalent to reality when designing any purposeful behavior system. In our furnace example, suppose that the furnace is imagined to be part of a large social program. An evaluation team might be sent out to analyze the reasons for success in the furnace program and to find out what makes it work. In the absence of real knowledge of how the furnace works, the evaluators might devise an evaluation plan that required them to synchronize their watches and to take simultaneous observations in both the living room and the other rooms of the house (including the basement). They would soon find the most obvious fact that, at the same time the temperature drops to its minimum in the living room, a noise starts in the basement. The noise continues during the period of rising temperature in the room. The noise stops at about the same time that the rise in
Figure 23. A Social Program
temperature in the room stops. The temperature in the room goes through a period of decay during which there is no noise, and then the cycle repeats itself. The correlations of the evaluators’ data would be very high (if their measurements are reasonably accurate), and after applying some complicated mathematical techniques, they might come to a very important conclusion: the noise should be tape recorded, hi-fi fidelity equipment should be purchased (considerably cheaper than furnaces) and installed in the homes of poor people throughout the country, and recordings of the noise should be played back loudly during cold weather. No one familiar with the structure and process of heating or of furnaces (for instance, a furnace repairman) would ever arrive at this conclusion. If the analysis had been kept in the language of the intervention being made—the operation of a furnace in a home heating system—such a mistake would be virtually impossible. Yet real-world analogies in social programs, the economy, and various forms of regulation are upon us and cannot be avoided.

The furnace example has some analogies in the program world, and figure 2-3 is a first attempt at describing a modern governmental program in the same way. In figure 2-3, we have replaced each domain with the analogous parts of a social program. The figure shows how a social program fits a skeleton diagram comparable to the skeleton of a home heating system. In subsequent chapters, the flesh is added to the bones. Again, only one feedback loop is shown where, in practice, many will exist.

Notes

During the past fifteen to twenty years, the terms *models* and *measurement* have come to mean many different things to many different people. In Part II, we describe what we mean by the terms.

The modeling discussion presented in chapter 3 is an introduction to some useful techniques for reducing reality to manageable proportions. The chapter distinguishes among several types of models used to provide the framework in which expectations are defined and measurements located.

Chapter 4 deals with measurement. By measurement, we include only certain well-defined methods of converting characteristics of reality to a number or a category. The discussion covers the kind and importance of measurement errors and emphasizes the interactive nature of model building and measurement—with the models providing the framework for questions the measurements providing the basis for altering the models.

Over the last ten years, we have found that three kinds of models are most useful for organizational work. *Logic* models, which present simple if/then sequences, are helpful in communicating the nature and purpose of a program. They are particularly good as a means of orienting the evaluation team and for making broad-brush presentations. They are less useful for analysis purposes since they cannot be systematically used to analyze cause and effect.

*Functional* models are the basic working models. These models, composed of traces and functions, graphically describe the interrelationships within the organization and between the organization and its environment and preserve cause-and-effect relationships, feedback loops, and significant patterns. They are the bedrock of the analytic effort.

*Measurement* models are anchored to the functional models and identify the measurements that can—or should—be taken in order to supply those in charge with the information they need to direct the activities of the purposeful organization.
This Page Intentionally Left Blank
Representing Complex Social Systems and Organizations

Has crime been reduced in the city through the use of a drug treatment program? Has unemployment been reduced through job training and public-service employment? Did the new recruitment practices improve the quality of the volunteer army? Did technique X improve the education of children in the schools? How much? These are common questions for evaluators, and reviews of the literature show that many evaluators, researchers, and policy analysts trip and fall over such questions simply because they fail to deal with the complexities of the social structures involved.

In chapter 4, we show that careful execution of several detailed steps is necessary in order to take even a single measurement accurately and reliably. The purpose of taking that single measurement is to reduce some characteristic of reality to a number or a category for manipulation and study. Unfortunately, problems involving governmental organizations can seldom be solved by the use of a single measurement. Sets of related measurements, representing the behavior of some organization or pattern, are usually necessary. This brings the evaluator quickly up against the problem of preserving for study the patterns and interrelationships of the measurement in time, space, dependency, cause and effect, and sequence—in other words, understanding a complex social system. The analyst must be aware of the relationships among measurements in order not to mistake what any single measurement actually represents. This presents the problem of capturing and describing structure, or patterns of interrelationships and flows within activities of interest, and the location of the measurement points within the described structure.

The part of reality to be captured and described depends upon the uses to be made of the information produced. The activity itself provides a rich and often confusing melange of potential information. An analysis of the activity, however, usually reveals a flow (or flows) of overriding interest—an essential factor(s) without which the entire enterprise cannot engage in the purposeful activity under investigation. In a business, for example, one such flow is cash. As we show later in this chapter, in a home heating system, an essential flow is energy in transfer; for a sanitation department, the key flow is garbage. Identification of the essential flow(s) permits the analyst to display the guts of the activity in the form of a functional model—a flow diagram in which the important interrelationships and functions are laid out.
Which interrelationships are important and which measurements will be useful depend partly on the expectations and intent of those in charge. The expectations are laid out in the form of a logic model—a diagrammatic representation of the if/then assumptions held by those in charge. In this chapter, we concentrate, however, on functional models drawn from the direct-intervention domain. Because this modeling activity is, by and large, shaped by the expectations that those in charge have for the program, we also touch on the descriptions of expectations and the logic models produced therefrom.

The term functional model is used to describe a representation that displays the characteristics of reality necessary for the use being made. Our belief is that flow diagrams and accompanying functional descriptions are usually the most useful techniques. In our experience, most governmental operations of interest (especially those with extensive feedback) are simply not captured even by most of the well-developed closed-form mathematical models or by simple logic models, although particular portions of a problem may be modeled in this way. Much social-science research, which gathers extensive data and then searches for correlations in it (often using multiple regressions as a model), might be characterized as measuring with the expectation that out of the measurements (through analysis) will come an understanding of (or at least clues to) the complex structure or process.

The approach espoused here could be characterized as modeling in an attempt to understand the structure or process at hand in order to decide what to measure and what analyses will be valid. Both approaches have their appropriate applications. Since further development of this matter is beyond the scope of this book, this chapter is aimed at creating an awareness of some of the concepts involved. The authors are available for debates.

**Finding the Direct Intervention: Where Functional Models Begin**

For the evaluator, quickly distinguishing the direct-intervention domain from that of those in charge is important. Those in charge, by definition, base many of their actions on abstract models of reality. These need not be mathematical models, in fact, they rarely are. The models are usually some kind of mental pictures, articulated in words, that represent what those in charge think reality looks like. The models may be accurate or inaccurate. What is important to realize is that all models leave out certain facts about reality. In that sense, every model of a real system is, as Ashby has pointed out, second rate. ‘People carry models around in their heads only because reality is too big, too complex, and too corporeal to be filed away in the brain and referenced as needed.

Direct intervenors, however, are in constant contact with
intervention reality—indeed, their activities are the reality of the intervention. While the direct intervenors also make models of the real process, they are at the place where the intervention exists, and their models usually have some basis in reality. For instance, the picture that any individual teacher has of what goes on in a classroom may be somewhat different from the reality of that same situation. The superintendent’s picture, however, is likely to be startlingly different.

Since a primary function of evaluation is to measure reality and to compare it to a standard, it is imperative that the evaluator locate the direct intervention quickly and plan some measurements to be obtained there around a model equivalent to reality. Evaluation results may be egregiously wrong if the evaluator mistakes someone’s rhetorical model for reality and inadvertently designs measures based on an inaccurate model of the activities and functions being carried out. This can easily happen, for instance, if data is gathered from reporting forms that do not report an accurate measure of a characteristic of the direct intervention. If the models are quite different from reality, the analyst might miss the fact that the characteristics selected do not even exist at the point of intervention.

In order to isolate the activities of the direct intervention from those that take place in the other domains, one question can serve as a test: If these activities were to stop, is it conceivable that the end product could still be developed? Product here is used loosely. For example, in a school system, the intended end product is an educated child, and the direct intervention encompasses the mutual activities of the teachers and the students. In a social-service program, the direct intervention normally includes a transaction between a government representative and a private citizen. In other types of programs (for example, a military training program), both parties to the transaction may be government employees or representatives. In all cases, however, the activities are central to the development of the end product.

For the most part, this book describes the preparation leading up to the evaluation of the impact of activities that take place in the direct-intervention domain. There will be occasions, however, when certain activities that take place in the other domains are to be evaluated. Much of the work of the General Accounting Office (GAO), for instance, involves the investigation of management. In those cases, it is helpful to isolate a group of activities to use as a surrogate for the direct intervention. The following example illustrates the point.

The Institute for Computer Sciences and Technology is the organization responsible for developing the standards that control the purchase and use of automatic data-processing (ADP) equipment in the federal government. As such, the organization is part of the federal government’s those-in-charge domain, carrying out an ancillary activity (see figure 1-5) that produces a contribution to the process (see figure 1-4) of automatically changing data into information. Since people often
want to know (with good reason) whether such ancillary activities are being carried out as expected (for example, Are the potential standards selected and developed through processes that plausibly will result in their beneficial use?), these activities are often targets for evaluation. In this case, the evaluators should isolate the set of activities within the institute that directly leads to the production of standards and should treat those activities as if they were within a direct-intervention domain. The inputs to that surrogate domain would be those things directly affected by the intervention—that is, technical information and models of how ADP centers work. The surrogate intervention itself is the process of configuring the technical information into standards that the intervenors believe can be applied in the ADP centers. (Note that the outcome of the evaluation of the ancillary activities will ultimately hinge on how accurate the surrogate-intervenor models really are.) In this case, the end product is a standard and the test question still applies.

If the interest is not in the workings of the ancillary activity but in whether it is helping or hindering the process of changing data into information, then the evaluators have to move to the places where information is the end product.

There are many ways to model the direct-intervention domain. We reiterate that the degree of detail chosen and the measurements selected depend on the use that those in charge intend for the evaluation. Another look at our home heating system provides some useful analogies.

**Drawing an Example from the Home Heating System**

Consider again a home heating system, oil fired with hot-water radiators. Many different models might be constructed to describe it. Reality, of course, is the heating system itself, a complex arrangement of tubes, pipes, tanks, heat exchangers, burner, and controls. Blueprints and a layout of connections would be one way of constructing a detailed model of this reality.

We start, however, at a much simpler level of detail to show how different models may be used to illustrate different viewpoints and different levels of understanding about various parts of the system.

*The Mystery Model*

This section illustrates the model of the home heating system that many of us have. Figure 3-1 shows two people in their living room. Their model of the heating system encompasses only a small part of the direct-intervention domain—namely, the radiator. Their view of the entire heating system is from the perspective of a simple goal-attainment model—that is, does the output of that thing keep them comfortable? Once they let the thermostat know their expectations, the administration
is on its own. The two people know that the thermostat transmits its directives to a furnace, but they do not know how; they know that the furnace makes the radiator get hot, but they do not know how; it is all a mystery to them. What they do know is that, if something goes wrong, they can call and talk to people at a certain telephone number and someone will come and fix the heating system. Beyond that, heating system is another phrase for mystery. The things impinging directly on the people in the room encompass the only part of the intervention that they understand. In addition, it is the only part of the intervention that they need to understand and all that has been modeled in figure 3-1. The structure of their organization for dealing with temperature-control problems is relatively simple.

Even at the mystery-model level, however, some important decisions must be made. The measurement involved takes place at a point in the room. That measurement indicates whether the level of the room temperature is comfortable. The first decision to be made involves an agreement as to what is a comfortable level. Once comfortable is defined, some decision must be made as to what is comfortable enough.

Thermostats sense temperature and control heat so that the temperature stays within a given range of a particular level—that is, when the temperature of the room falls a certain amount below the optimally comfortable level, the furnace is turned on, and when the room temperature rises a certain amount above the optimal level, the furnace is turned off. Of course, disagreement on the level of the temperature is possible. Also, the allowed range of variation about the level may be so wide that the room is not comfortable enough part of the time.

Therefore, the second decision that must be made is: How accurate must the measurement be? These decisions are by no means trivial. Extrapolating to the world of policymakers, the Federal Communications Commission has, as one of its mandates, the setting of picture-quality standards for broadcast and cable television. Particularly in the area of cable, the argument over what is good enough and how it is measured had still not been settled as of this writing. In the area of federal day-care policy, the Office of Child Development was responsible for setting standards for adequate day care and for minimum acceptable vari-
ations from these standards. The Office of the Secretary of HEW, the Congress, and on occasion, the president of the United States have engaged in the argument over what is adequate enough.

Once those two decisions are made, then the measurements can be taken and the real question answered: Is the system working?

A rudimentary logic model, shown in figure 3-2, can be extrapolated from the mystery model. As the logic model indicates, the only question the homeowners are interested in testing is: If the thermostat is set correctly, then are we comfortable? The mystery model is adequate for this purpose even though almost the entire direct-intervention domain is beyond its view. If the system fails, then an alarm is given and someone else (in this case the furnace company) diagnoses the problem using a more-complex model. The measurement points for this question are easy to locate, and the model is adequate for the purposes of the initial evaluation.

In this case, the evaluators are aware only of their immediate surroundings. The furnace, the rest of the house, the furnace company, and the remainder of the world are all mysteries. There are other ways, however, of constructing models, and there are certainly more activities that could be included in the model for other uses.

The mystery model is one of a class of models that engineers call black boxes. (Economists usually call them input/output models.) These models depend only on capturing the flows across a closed boundary. The principle is that a fence virtually encloses an entire activity (in the mystery model, the activity is everything to do with the heating except the radiator and the thermostat). The evaluators do not know what happens inside of the boundary (they cannot see into the black box), but they can find out all of the things that go in and come out as long as those things are recorded in the area outside the black box. The things designated as beyond the boundary of interest depend on who is interested. Office personnel at the furnace company would have a different perspective than the people inside the room.
Suppose the people in our mystery model determined that the house was not comfortable enough. They would activate the alarm links between the house and the furnace company. Someone in the office would take the call and would, if common sense prevailed, ask the callers if they had checked the thermostat. In other words, the most basic question is answered first. (This is true whether the evaluation is of a furnace or a manned space flight. There is no point in evaluating the flight performance of complex instrumentation if the rocket never got off the ground.) Once they determine that the thermostat is indeed set correctly, office personnel can then proceed with an evaluation using a different mystery model, or black box. This might be a model that views both the furnace company and the house as black boxes. Figure 3-3 illustrates that mystery model.

An additional question that is asked here is, of course: What are the interrelationships between inputs and outputs; what do they cause or do? The furnace company is putting oil into the house; it is also putting bills into it. The people in the house are sending money back to the furnace company. An alarm arrangement of some kind assures that, when the people who live in the house feel that the room is not comfortable enough, they can alert the furnace company. Look at your home heating operation that way, and most of the normal inputs and outputs for your heating system would be captured. (Heat loss and efficiency would introduce another level of analysis and call for more-complex models and information collection.)

If you look at the assumptions involved, they might be as simple as those in figure 3-4. The oil company is putting in oil and putting in bills. The bills seem to produce money, and the oil seems to produce an absence of alarms. These linkages, or interrelationships, exist over time. In this black-box model, the measurements made are predicated on a simple logic model that implies these interrelationships. Notice that the pattern of furnace activity has not been included.

To use this logic model, the office personnel would check the files to make sure the bills had been paid. If not, they would check to see if the
bills had been sent out. If they had, they would check to see if the oil had actually been delivered. If not, the problem has been identified. If it had, more extensive evaluation is required and a different model must be constructed. Office personnel alert the service department that then sets off to work with the help of a functional model. Thus, measurements from various points on the model feed into a series of sequential decisions. If the solution is not obtained, we exit from this model.

The Functional Model

A Home Heating System and Energy Transfer. Looking a little more carefully at what happens as different models are considered, we see that the furnace company actually uses at least two models (figure 3-5). The black-box model is normally used until investigation indicates that the problem is not one of the inputs or outputs. Something inside the last black box described requires servicing. Then the furnace company must switch to a different model—often a functional model. As implied earlier, a functional model describes how something works. It is important to realize that the knowledge of how a thing works is often necessary to test the plausibility of the logic model. For instance, the logic statement, “If I call their number, someone will fix my heating system,” is implausible if the furnace company has gone out of business.

It is irrelevant whether the home owner or office personnel knows anything about the functional model (as long as a service person is available). However, the person who services the furnace should know a great deal about the functional model—that is, all the steps including the
control circuits that lead from the presence of oil in the tank to a room that is comfortable enough. This is because the service person has a use for a more-detailed model in carrying out work. Figure 3-6 displays a model of the process (omitting the control circuit for simplicity).

In constructing a functional model, an essential flow (or flows) of interest must be identified, some flow that serves to unify the system and that is sufficiently descriptive to permit all parts of the relevant reality to be modeled around it. The unifying flow of the functional model of a furnace is energy, for whose transfer and transformation a well-developed body of knowledge exists. In many public programs, finding the proper essential flow for the model and an accompanying body of knowledge is considerably tougher. Without it, however, much evaluation is performed blindly and poorly. Development of these unifying concepts and accompanying models is one of the state-of-the-art problems of our day.

While we cannot furnish an algorithm that spews out appropriate essential flows, we can offer a few heuristics. First, as we mentioned earlier, the flow(s) should be essential to fulfilling the purpose of the activity. Second, that flow should be expected to wend its way through the salient parts of the organizational entity. In a home heating system, such an essential flow is energy as it is transferred from one place or state to another. This essential flow is unifying because it flows through the system—unlike oil, for instance, that, while essential, does not stray far from the oil tank without being burned and ceasing to be oil.

Using a unifying essential flow means that when a map is made of an activity, or a model developed that describes it, that something can be traced through in detail to determine what happens at each point of interest. The structural relationships revealed by the trace can be displayed in a model that is good enough for the uses that are going to be made of it. In the case of the furnace, using energy as the unifying flow, the functional model looks more or less like figure 3-6. Starting at the edge of the house—where the black box started—the first step is to put in potential energy in the form of oil. Oil being available, the next step is burning the oil’s potential energy to produce heat. The heat is then transferred to the hot water, which is another way of carrying the energy. Circulating the hot water is a way of taking the energy out to the radiator. When hot water gets to the radiator, then the radiator (which is another heat exchanger) moves the energy from the hot water to the air around the radiator. Room-air circulation is the method of moving the hot air around in the room to achieve a comfortable temperature. If a proper temperature is maintained in the room, then the room should be comfortable enough.

So far, this has all been a functional diagram of what happens in the energy-transfer process. It displays the structure and relationships of the process. This functional diagram can also become the device for choosing, defining, and relating measurements. It is a proper base for locating and describing many measurement points. When the control
system has been added, the diagram displays most important structural and operational aspects of the system is enough detail to start making measurements aimed at evaluating the system (in this case evaluating what is wrong). If a problem occurs in the burner motor, a new model may be needed that describes burners and their measurement in more detail.

There are two ways of immediate interest to measure the temperature of the room. Using the simplest way, the thermostat (and the owners) sense the temperature in the room (outcome measures). That measurement was illustrated in the mystery model (figure 3—1). The other way is for various points in the process to be measured (process measures).

Figure 3-6 shows examples of some additional measurements that could be made each time the energy is transferred to a different form. This illustrates how a measurement point can be indicated on a functional diagram and a measurement shown. The questions that this model answers are: What are the steps of the process in terms of energy transfer? Where could measurement be taken to see what steps are being performed properly? What could be measured at the various points? Again, note that before going to a more-complex model, the evaluator answers the more-basic questions. As detailed descriptions of measurement points and measurements to be taken are added to the functional model, a new model, a measurement model, is created. Measurements are discussed in the next chapter.

We have used a home heating system to illustrate how models are created from different perspectives. Note that the functional model shown assumes that all of the questions that the office personnel (administrators) asked were answered affirmatively. If it had turned out that one of the administrators’ questions—for example, Did the oil get to the house? had been answered with a no, the functional model would look quite different—more like a route map from where the oil was stored to the house in question. Many different functional models can be used in developing, defining, and explaining measurement points, measures, and measurements; the selection of the appropriate model depends on the intended use of the information to be obtained. The models discussed are shown arrayed in figure 3-7, along with two other possible models. Figure 3—7(d) represents the complete manufacturing drawings and assembly and operating instructions for the heating system. Figure 3-7(e) is a picture of the heating system itself in actual operation. Of course, many options exist for performing actual measurement. For instance, the temperature of the hot water could be measured by a thermocouple placed in the water or by a gauge that measures pipe expansion. Which measures to make depends upon one’s purpose for measurement, time and money, equipment, skills, and of course, what you think might be the problem.

Notice that in all of the examples in figure 3-7 four important things are repeated:

*Patterns:* The models display patterns and interrelationships

*Measurement points:* The models become road maps for locating measurement points relative to everything else.
Figure 3-7. Five Levels of Models of a Heating System
**Measures:** At most measurement points, many measures are possible.

**Uses:** The use to be made dictates how much of reality is represented, what points are selected for examination, and what measures are taken from these points.

In each case, the model is equivalent to the reality of importance to the people who have questions that need to be answered. These models are fundamentally and philosophically different from the logic models (figures 3-2 and 3-4) that are based on beliefs about reality. The distinction is discussed in detail in Part III.

**A Garbage-Collection System and Garbage Transfer.** The primary direct intervention or service in a garbage-collection department is picking up the garbage and disposing of it. The provision and servicing of trucks, financial planning, establishing routes, and so on are ancillary activities that have their payoff in effectively improving the direct intervention—namely, the collection and disposal of garbage. Those who collect and dispose of the garbage are the major active participants in the direct intervention. Their activities and effects must be examined as a starting point for developing functional models of the direct intervention of a garbage-collection agency.

In drawing a functional model of a garbage-collection system, the unifying flow would probably be garbage transfer, and thus garbage would be traced. A functional model that looks like figure 3-8 might be involved. As this model shows, garbage is deposited in and about the streets (presumably in cans). A truck then picks it up and takes it to the disposal site where it is disposed of. At this point, the garbage is all gone as far as this evaluation is concerned. The unifying flow has been tracked all the way through from the time people put the garbage on the street to the time that the garbage collectors made it disappear. This functional model is somewhat easier to create than that of the home heating system because the essential flow used in the model (garbage) does not change state (except possibly to decay a little) until it is disposed of. Many different measurements could be made from this functional diagram. If, however, a set of measurements is defined in terms of garbage, a simple model of stocks (at different locations) and flows can be developed. (Notice that in the heating system simple stocks and flows were of energy but that some physical transformations had to be represented at each box.)

The selection and definition of particular measurements—how they will be performed, where they will be taken, and how they will (or can) be used—is crucial to the design of the evaluation. For now, we discuss them only to illustrate how they relate to the functional model itself.

As we showed when we discussed the furnace examples, different people on different occasions have different questions about the operation of a process. The questions asked determine what kind of model will be used and where the measurements will be made. The administrators of the furnace company were trying
Figure 3-8. Garbage Transfer
to answer the question: Why is the room not comfortable enough? Measurements were made first at the points at which the question was most likely to be answered easily. When a different question was asked—for example, Why is the furnace not working?—different measurement points had to be chosen, and a different model was required to illustrate them.

What kinds of questions are typically asked about a garbage-collection system? Often, the question is: How much garbage is being collected? The answer to that question can usually be found at the garbage dump—the right-hand side of figure 3—8. Or the question may be more specific: How much garbage are we taking from homes, from playgrounds, and so forth? This measurement (perhaps in ton miles) would probably be made at the middle of the diagram because now the loads of individual trucks are measured rather than the aggregated garbage. Note that these measurements say nothing about the adequacy of the service vis a vis the cleanliness of the streets. For all the evaluator knows, an equal amount of garbage may still be strewn about town.

If the questions asked are: How much garbage is on the streets? and What conditions are the streets in?, then the evaluator has to look at the streets, making measurements at the far left of figure 3-8. The Program Evaluation Group of the Urban Institute designed an evaluation to answer those questions for Washington, D.C.² This evaluation measured the cleanliness of the streets by comparing them with standard sets of photographs. By measuring cleanliness of streets, it is possible to feed the information to the supervisors of garbage collection and affect the patterns of service used and the types of service given.

The garbage example illustrates that relevant measures are not always at the end of the functional chain. It really is a matter of what question is being asked. Where and what to measure is conditioned by what uses are to be made of the information, to whom you wish to give the information, and what is to be affected. Different measurement models can be constructed from the functional model to address different types of questions. The functional model itself, however, should be tied directly to the reality under study. The uses to be made of the information simply indicate what parts of the model may have to be further developed and in what detail. The measurement model is developed in order to keep track of the location and interrelationship of all the measurements. The relationships in the measurement model come from the functional model.

A Local School System and Knowledge Transfer.

In a local school system, the activities that survive the test question about direct interventions are those that involve teachers spending time with children. The attempt to alter skills, knowledge level, and
capability through the teacher/child interface is the major direct intervention made, and this systemwide direct intervention is the work that must be examined, modeled, defined, and measured in major assessments of school systems.

Compare this activity with at least two other major kinds of activities that usually exist in the those-in-charge domain of a school system. First, there are ancillary activities intended to either enable or to enhance the primary direct intervention. Those activities include providing buildings and supplies, the work of operating personnel and payroll departments, certifying and training teachers, and some of the activities of curriculum and instruction departments. For any of these ancillary activities to actually affect the education of children throughout the school system, they must somehow affect the intervention that takes place in teacher/child activity. The direct intervention (teacher/child) becomes the eventual testing point for the results of any ancillary activity—that is, if people want to know whether the county curriculum program works, they will in the end have to assess its effect upon the education of children by teachers in classrooms. (The step that evaluators from the Institute for Computer Sciences and Technology took when they went out to model an ADP center was to use data transfer as the essential flow.)

The second kind of major activity that usually exists—in-house activity—may have only a tenuous relationship with the direct intervention, but usually no direct connection at all. Such activities might include writing speeches for delivery to meetings of educational administrators. If the teacher/child is taken as the point of direct intervention, then most work that does not reach or affect the bulk of the teachers in the system would be regarded as in-house, including some that might at first appear to be ancillary.

As we define it, only the teachers and those who immediately supervise them (plus a few other people who spend the bulk of their time with children) fall into the direct-intervention domain in a school system. That is where the functional modeling must begin.

So far, the models shown have been relatively noncontroversial. Everybody agrees that energy transfer is the unifying concept that runs through a normal home heating system. No one would seriously dispute that garbage is what garbage collection is all about. Unfortunately, universally accepted unifying flows are not available for most evaluations. The evaluator often has to develop one. In this example—the Program Evaluation Group’s evaluation of the Atlanta Public School System—the unifying flow chosen was the transfer of identifiable, unambiguous knowledge (for example, $2 + 2=4$)

At first, the Program Evaluation Group could do no better than a black-box model. All that was really observed was that children went into
the classrooms, spent time with other children and teachers, and came out again. Other people were in the school, but the evaluators did not know what they did. Other people were also in other buildings in the school system, but again, it was not clear what they did either. There were area superintendents, a county school organization, and a school board. The team was a little baffled by all the different and sometimes conflicting things that people said they did compared to what they actually seemed to be doing. In fact, this is what a school system looks like to many parents.

The hard, confirmable facts seemed to be that children went into the school room, children came out of the school room, and they were changed. At least, everyone assumed that they were changed. There were some measuring instruments to test that assumption. If indeed the children were changed, it was possible that the change could be measured and reported. In other words, the black-box model (figure 3—9) was sufficient to answer the question: Does it work?

Obviously, if any questions other than Does it work? were to be answered, a functional model had to be developed. Choosing knowledge as the unifying concept and looking at knowledge transfer, a typical model looks like figure 3-10.

Various logic models, such as figure 3—11, could be extrapolated from the beliefs held in the system and examined against the functional model of figure 3—10. First, if the teachers have an accredited degree, then they are qualified to teach. (After a period of observation, teach was defined as children and teachers spending time together in a classroom. That somewhat vague definition was used because it was the only one that encompassed the plethora of methods and resources used.) Second, if they are qualified and have certain resources, then teachers can and will teach. Third, if the qualified teachers (with

![Figure 3-9. Mystery Model of a School System]]}
resources) spend a given amount of time with children, then children will learn. Fourth, if children learn, then their scores will be good enough on their tests.

Each of the four assumptions in the logic model is testable against the reality of the direct intervention. On the most basic level is an existence test— for example, Do the teachers in fact have the required academic degrees? Have they been given resources? Do they spend a given amount of time with children? This is analogous to seeing if the thermostat has been set right. At the next level is a check for noise in the system—for example, Are Tuesday’s lessons being drowned out by the practice sessions of the school orchestra? This is analogous to seeing if a cold draft is blowing on the thermostat. If the existence proofs are there, and if no noise is contaminating the system, then more-detailed functional models are required in order to lay out some of the subtler questions implicit in the logic model—for example, Are the questions on the children’s tests related to what the teachers are teaching in the classroom?
As each assumption link in the logic chain is compared to its plausible, operating counterpart in the direct intervention, then—but only then—can we conclude that one or more of the logical assumptions is wrong. This is analogous to checking each part of a home heating system before concluding that the engineering design is faulty. In reality, the measurements suggested by the logic required consideration of many other intervening variables.

**Why Make a Model of the Direct Intervention?**

Once the logic model (usually obtained from those in charge) has scoped and displayed the areas of interest, the evaluators are well advised to model the direct-intervention domain as the next order of business.

Detailed knowledge of the exact process of the direct intervention will often prevent the evaluators from being badly fooled about what is measurable and, perhaps, from making silly measurements or trying to make impossible ones. Chapter 2 contained an illustration of the way in which perfectly sincere evaluators might be led to believe that the noise from a furnace heats the house. In fact, an expensive set of measurements might have led to the inference. Indeed, some consequential policy decisions might have resulted from that evaluation. Either everyone would have been very embarrassed or the target group for the program would have been very cold. One reason we chose the furnace analogy is that it is nontraumatic. Particular examples from, for example, health, education, welfare, labor, law enforcement, energy, the Council of Economic Advisors, and so forth would have been apt to traumatize the erstwhile owners of the example and, at best, to lead to lengthy exchanges of correspondence for the authors. Our experience with many government agencies indicates that a detailed understanding of the nature of the direct intervention will often protect against spending too much time and money on unanswerable questions or mistaking measurement of one thing for evidence of another.
Usually, this type of error can be avoided if the following steps are taken:

Determine the important direct interventions and their locations.

Go to those locations and determine the following:

- The actual existence of the described direct intervention;
- The exact observable pattern and nature of operation (make a functional model or flow diagram of it);
- The major inputs and the enabling, or ancillary, interventions affecting them;
- The measurable (or potentially measurable) outcomes.

Having made these determinations about a direct intervention, it is now easier to trace the linkages in a similar way (or not to trace, if there is no intervention or no linkages) through the those-in-charge domain of the organization to the direct intervention and to come to (at least preliminary) conclusions about the kinds of information that might be usefully provided to different levels of those in charge and the mechanisms through which decisions made by those in charge affect (or fail to affect) direct-intervention activities.

Our experience has been that construction of a functional model equivalent (or even possibly equivalent) to the direct intervention by the evaluator greatly facilitates much of the subsequent evaluation activities. With the functional model in hand, it is relatively easy to visualize the types of models that are needed to capture both the relevant management activity and the expectations about intended outcomes and impacts of the direct intervention.

Purposeful behavior operates through a completed feedback loop(s). However, policy questions are usually framed in terms of: If I do this, will I get this effect? Thus, the evaluation designer is usually faced with constructing a working model of a situation that can be affected strongly by three sets of activities:

1. Activities in the those-in-charge domain that may influence the direct intervention;
2. Activities in the direct-intervention domain that affect what is likely to happen
3. Activities through which the direct intervention (if successful in producing desired immediate outcomes) is thought to influence or create the desired impacts.

Note the critical position of the direct-intervention domain. It is the focus
of activities in the those-in-charge domain and is thought to be the cause of impacts. Without knowing the patterns of what is really occurring at the point of intervention, there is no way of knowing which activities performed by those in charge are relevant or what behavior might conceivably be attributed to the intervention.

While we have emphasized the functional model as the basic tool for analysis, the three models mentioned (logic, functional, and measurement) are all necessary tools, and special attention should be paid to the relationship of the different models to each other, to expectations, and to reality. Samples of the three models from one of our early studies are shown in figure 3-12.4

Figure 3-12(a) represents two of the many expectations held for this child health program. Figure 3-12(c) depicts children flowing through one of the projects in the field. Figure 3—12(b) shows a set of actual measurements. Notice that, while these models are very different from one another, they are all concerned with aspects of the direct intervention.

Government programs and their descriptions are often keyed around enabling interventions—that is, the Department of Education (or the local school system) gives funds for training teachers in a special technique; the Law Enforcement Assistance, Administration (LEAA) funds better radio-dispatch systems for police; state planners are funded in a variety of programs; the National Institute of the Mental Health (NIMH) funds technical-assistance programs for community mental-health centers; or the national energy program places a stiff tax on gasoline. The language used to describe such enabling interventions is often in terms of some desirable direct result that is supposed to happen. Sometimes the discussion is even in terms of the social impact or benefit. The evaluator will often be faced with questions like: Is everything working better now? and Has this solved the problem?

Where to measure? Success at the ubiquitous actual point of direct intervention ultimately determines whether the entire string of enabling actions was a success. Ergo, examine the direct intervention and its effects first (or at least early) when designing measurements or scoping a problem.
(See following page for conclusion of Figure 3-12).
(c) Functional Model of a Particular Project

Figure 3.12. Three Types of Models: Logic, Measurement, and Functional
Notes


The recurring theme that ran through Part I was that evaluation, properly, is a process of measurement, comparison, analysis, and use. In Part II, a major point was that because measurements must be made in the direct-intervention domain, a considerable portion of the evaluator’s time must be devoted to that domain because that is where the bulk of the modeling activities is focused.

Note, however, that the standards for comparison are not developed in the direct-intervention domain but in the domain of those in charge. In addition, much of the use made of evaluation will be made by those in charge. In other words, evaluation activities should begin and end with those in charge. They begin with the development of questions, and they end with the use of the answers. That is the framework in which the evaluators must work.

Viewed in that perspective, evaluators clearly should know a great deal about those in charge—for example, how closely their conception of the program corresponds to the program as it exists; by whom, and under what conditions, evaluation results can be used; the comparison standards likely to be accepted; and the real expectations for the program, as distinguished from hopes, dreams, desires, and posturing. This leads to a much different orientation than that of most books about evaluation, in which the advice about dealing with those in charge can usually be boiled down to two steps—that is, (1) find out what the decision maker’s questions are and (2) answer the questions. Unfortunately, evaluators tend to fall on their faces when attempting to take those steps. One reason for this is the existence of multiple organizations, levels of organization, and forces involved. A second reason is the paucity (or even the absence) of decision makers. (Tom Kelly, of the Environmental Protection Agency (EPA), maintains that there are many more position takers in government than decision makers.) A third reason, and one we touched on earlier, is that many of the questions initially asked are simply unanswerable. In a sense, people who deal only with policy and not with process seem to assume that once they have decided on their policy—whether it is better mental health, lower crime, better education, lower unemployment and inflation, and so on—that it is simply a matter of turning the thermostat and the change will occur. Many policymakers and analysts often seem to be turning thermostats that are not connected to furnaces or houses.

If the goal of implementing purposeful behavior through evaluation is to be reached, much information from those in charge is required for the design of the evaluation and for creating conditions that lead to the use of the information.
eventually produced. We therefore devote this entire part to an examination of those in charge—and at that, we barely scratch the surface.

Chapter 5 expands on our earlier (chapter 1) description of the domain and highlights some of the problems in coping with it, particularly the problem of drifting semantics. In chapter 6 we explore some of the factors that influence the acceptability of given measures and comparisons to those in charge. Chapter 7 presents some heuristics for threading the maze of those in charge. Seven clues are discussed: two of them are aids to finding the right people to talk with; the other five involve the things that are talked about. Chapter 8 reexamines purposeful behavior from the perspective of uncovering the purpose. The distributed-incentive structures of large governmental organizations are examined.
Recapitulating some of the material from chapter 1, the those-in-charge domain lies between the source-of-authority and the direct-intervention domains. Those in charge translate the language and intent of the enabling intervention into directives and guidelines for the direct intervenors, pass the money along, and are in charge of seeing that everything operates smoothly. The those-in-charge domain contains some people who do not deal with or influence in any way the direct-intervention domain (for example, the Secretary of Interior’s office manager). The those-in-charge domain also contains some people who deal with the direct-intervention domain but who have no authority over it (for example, the people responsible for procurement of park-service uniforms). Finally and most important, the those-in-charge domain contains people who can and often do influence the activity in the direct-intervention domain. The influence exerted by these people depends, in part, on their rhetorical models of the direct intervention—that is, on their idea of what is actually happening there. In response to its own purposes and experience, each level of management will have its rhetorical model of the direct-intervention domain.

These rhetorical models often do not resemble the actual direct intervention in important particulars. In fact, the various rhetorical models held by those in charge may not resemble one another. A major problem facing the evaluator is the reconciliation of the important rhetorical models—the definition of words, phrases, and concepts so that those in charge at different levels in the organization agree on a common rhetorical model that means the same thing at all levels. This involves both identifying the important models and finding a common language. Importance can only be sensed by learning the structure of the organization, meeting the players, and using any information, rumor, and gossip that comes to hand in order to recognize actual operating loops. The search for a common language is greatly helped by the early construction of a family of testable models (models that can be tested against standards of plausibility and reality). (The families of models are dealt with at length in the Part IV.) The testable family includes three representations of the combined beliefs held by those in charge: (1) a logic model, described in chapter 3; (2) a testable functional model, which describes the beliefs of those in charge about the program in question; and (3) a testable measurement model, which displays the measurements necessary to answer the questions of those in charge without regard to
whether the reality of the direct intervention will permit the measurements to be made. As these testable models are developed and discussed with those in charge, disagreements and discrepancies in concepts and beliefs will surface and may ultimately lead to the reconciliation of the various definitions.

All of the models mentioned (and a few more) will be necessary to carry out an evaluability assessment and eventually to construct an evaluation design. It is best to start drawing them early in the investigation.

In addition to the models that are necessary to carry out future work, one further model often proves to be extremely useful—namely, a functional model based on observation of the relevant portions of the those-in-charge domain. While this model is not, strictly speaking, a requirement for evaluation design, it does help the evaluators sort out who can do what and to whom and saves a great deal of the time that is usually wasted talking to people whose functions are the equivalent of painting the furnace blue.

The Planners’ Dream of the Government Organization

For simplicity’s sake, much of the presentation in this book deals with a single loop of information—a furnace-like loop involving actual process, evaluation of results, information for those in charge, and actions to change the process. The very basic elements of evaluation and management are illustrated in each single loop involving actions, measurement of the results of those actions, and decisions about the continuing actions. A single loop of immediate feedback and management control occasionally occurs at the points of direct intervention into society, especially in decentralized organizations.

However, in exploring large local- and federal-government departments, it quickly becomes obvious that this single loop is inadequate to describe what really goes on. Because most large organizations have many feedback loops, all operating at once, modifications and elaborations must be made.

Typically, the elaboration offered by texts and by planners and policy analysts adds one superloop to the furnace model, a superloop that includes the planners and policy analysts. Figure 5-1 is such a prototype model.

The dream, described in much of the planning literature, is about how policy analysts use comparisons and analyses of the operating process to draw longer-term, higher-level insights than those needed by operational management simply to keep the process running.¹

The policy analysts receive agendas of issues and goals from the policy management. They then mesh these with factual information received from the field. They develop a slate of new initiatives (from which the policy management can then choose on the basis of goal statements, objectives, cost/benefit data, evaluation results, and from
mathematical projection). The policy management then chooses a new policy thrust for implementation by the people in the field. At this point the policy management turns to its planners. The planners translate the new policy initiative into basic terms that the operational management can understand. (There is a strong implication in the literature that the operational people are far less sophisticated than those in charge, so that much guidance and regulation is required.) Of course, this document produced by the planners is usually called The Plan.

This Plan may even contain additional information requirements to be collected in the field so that the analysts will be better able to determine the results of the new and better policy initiative. The Plan is, of course, so complete and detailed that it forms the basis for the subsequent activity of the operational management. The operating managers, in turn, alter their own process of intervention to bring it into line with each new plan. The continuous activity of measurement, comparisons, and analysis in time records the results of the new policy
initiative. These latest results are then used by policy analysts, management, and planners to develop new policies, to make choices, and to send out new activity instructions. These modifications are ground into a new plan by the planners and eventually arrive again at the operating process. To our knowledge, this dream has not yet been realized.

The Governmental Organization during Waking Hours

The planners’ dream is destroyed, in real life, by the fact that many formal and informal loops control the behavior of the different residents of the those-in-charge and the direct-intervention domains. These loops are connected to the Congress, the president, dominant lobbies, and a host of other influencing factors that exist in the environment of any large government agency.

A detailed description of those factors and their relationships to those in charge is well beyond the scope of this book. A simple analogy, drawn from our furnace example, will suffice as an illustration of the complications that arise.

Chapter 2 noted that the purpose of the furnace was to keep the house temperature at a comfortable level. Imagine, then, those in charge of the home heating system. Suppose that the family consists of Momma, who tends to get drowsy when the temperature exceeds 72°; Daddy, whose anemia causes his feet to get cold when the temperature drops below 70°; and the twins, who work up a sweat practicing kung fu in the recreation room. These are the people who define comfortable, and these are the people with access to the thermostat.

Now add some factors from the environment. Suppose that the Organization of Petroleum Exporting Countries (OPEC) decrees an embargo on oil to the United States. The furnace company is forced to ration oil for a period of time. Suppose also that the family is not economically self-sufficient but relies on a tightwad uncle to pay the fuel bills. The uncle is sufficiently sensitive to pressure from the rest of the family not to let his relatives freeze to death, but the fact is that he does not really care whether the house is comfortable. His main concern is that the fuel bills do not exceed a specified amount during any one month. You might also include a neighbor who is threatening EPA action if the chimney does not stop belching all that black smoke. Given that composition of those in charge, plus the environmental factors affecting behavior, the odds are very small that the operation of the furnace will ever attain a stable condition. The superloop of the planners’ dream is simply irrelevant. The most that evaluators can hope for is the ability to arrive at a definition of comfortable that Momma and the twins can all live with and a set of measures that is within the operating constraints yet reasonably consistent with the family’s needs and the requirements of the environment.

If any actions are to be taken by those in charge in response to information created by an evaluation, it is important for the evaluators to
get a firm understanding of what different groups among those in charge can use as models of the process, what the groups expect to learn about the process through the information created by the evaluation, and what constraints are imposed by influencing factors. The evaluators have to know what each group believes the intervention or service process to consist of and what each group among those in charge considers to be proper measures and comparisons to use in determining success.

Because there are often multiple levels of those in charge between the authority and the intervention, gathering this information often resembles hopping about between isolated islands and learning the language and belief structures of the group of natives living on each island. Often the abstractions and semantics used change perceptibly as geographic and/or organizational distances from authority or intervention change. What is acceptable as a measure or comparison at one level may often be unacceptable at another. Further, those in charge of governmental organizations may be much more polished and practiced at answering questions than the people in the direct-intervention domain. Developing an evaluation design means developing methods of clarifying what the questions and expectations of the different members of those in charge really are on the one hand and developing ways of answering them with information produced from measurement of actual intervention activities on the other. The material in the next section is an attempt to familiarize you with some conditions that are usually encountered in dealing with those in charge.

**Multiple Levels of Organization**

The first factor normally encountered is multiple, distinct levels of organization between the authority and the intervention. This is true of every organization of size. Many professionals can, in the course of their work, ignore the different levels without any great harm. The evaluator or evaluation designer can almost never do this. In organizing to produce and use information, the opinions, beliefs, needs, and power of different levels of the organization are likely to be critical in choosing measures of success, in determining the accuracy to which the answers must be measured, in deciding how and whether to proceed, and in obtaining any use of the information once it is produced.

Different levels of organization are illustrated by the examples in figure 5—2. The typical industrial organization has a board of directors and a charter that, along with a body of corporate law and regulation, provide the legal authority for the organization. Within the those-in-charge domain, any particular company might have someone like a president, executive vice-president, director of production, manager, and supervisor. (If the entire organization was shown, the organization would expand at the bottom like a pyramid due to multiple directors, with each director having multiple managers and each manager having multiple supervisors.)
Tracing this line of authority into the direct-intervention domain would lead to line supervisors and workers. The worker performs actual work on the products of the company, and the line supervisor directly oversees the workers. (This was the heuristic given for roughly drawing the boundary of the direct-intervention domain.) Lines of authority are identified by tracing the linkages between the worker and the organized authority and laying out what is found.
Those in Charge of Government Agencies

In a federal program such as the Community Mental Health Center (CHMC) Program, the direct intervention is usually accomplished by a caseworker or psychiatrist. Several years ago, when we attempted to trace out the supervision or policy chain in this case, we arrived at a structure such as that shown at figure 5-2(b). Congress had established funding for these centers by law (although each particular operation was also dependent on both local law and federal guidelines). In the those-in-charge domain of this program, we found the president of the United States; the secretary of HHS; the deputy secretary for Health; the Alcohol, Drugs, and Mental Health Administration (ADMHA); the NIMH; the director of the entire Community Mental Health Center Program (HHS regional offices could have been included); and the local director of a particular center. All of these players could affect CMHCs in varying amounts and ways. None was directly involved in the service or intervention.

Familiar variations on this theme can be found on the local level. School systems, for instance, usually look something like figure 5-2(c). Here the legislative authority is usually a combination of state and local laws and actions by a local school board. If the connection between this school board and a classroom teacher is traced, a superintendent, deputy or associate superintendent, and principal will be found be in the those-in-charge domain. The teacher and, perhaps, an assistant principal would lie in the direct-intervention domain. If the principal was the primary direct supervisor of teachers, then the principal would have been included in the direct-intervention domain as well.

The examples are not meant to be exact, but are simply given as illustrations of the type of mapping that is necessary in any given case in order to bring into focus the different levels and positions in the organizations where changes may be made that affect the actual work of the organization, where questions that are to be answered by the evaluation originate, and where discussions may be held about what type of information may be necessary and how accurate it must be.

The job of the evaluator is to try to produce a body of systematic information linkages—through measurement, comparison, and analysis—from the direct-intervention domain to the decision elements in order to establish a basis for influencing the decision process (at least in part) with soundly based performance information over a period of time. The furnace examples of chapter 2 provide simple analogies of such a process in operation, although in actual practice the operation is considerably more complicated than a single measurement and comparison loop. Methods of proceeding through such organizations are discussed in chapter 7. This chapter simply indicates the nature of the organizations.

Multiple Levels of Semantics

As the evaluator examines the organization to pin down the questions that must be answered, the groups that may take action, and the
types of actions they may take, different groups of Those in Charge—and especially different levels—often are found to be speaking slightly different languages. They are all speaking English, of course, but distinct semantic variations can be encountered that must be understood and interpreted if the resultant evaluation, when implemented, is to be useful in helping them to improve the direct-intervention process.

Part of the change in language represents the generalization of function that is the purpose of a multilevel organization. Another part of the language change, the part that usually proves troublesome for the evaluator, is a semantic drift away from specific referents toward global abstractions. As this occurs, program descriptions become progressively less tangible, and the questions for the evaluator become progressively less answerable.

Consider, for example, how several levels of an organization might describe a dog and what each level might want to know about the animal. At one level it might be referred to as that dog, implying a set of characteristics particular to the dog at hand (for example, color, size, sex, breed). At the next level, functional description, attesting to certain expectations of the dog’s behavior might occur. At still another level, generic policy terminology might appear; and it would not be surprising to find (especially at budget time for the dog program) the mutt described as man’s best friend.

Table 5-1 illustrates the semantic drift and shows the kinds of questions associated with the different organizational levels. As the table

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-level policy</td>
<td>Man’s best friend</td>
<td>What is the level of friendship in the United States?</td>
</tr>
<tr>
<td>Policy</td>
<td>Director of nutrition programs, keeper of peace</td>
<td>To what extent are sustenance and tranquility related to friendship?</td>
</tr>
<tr>
<td>General functional</td>
<td>Hunter and protector</td>
<td>Is adequate game-trailing education provided?</td>
</tr>
<tr>
<td>Direct description</td>
<td>That dog</td>
<td>Does my dog have fleas?</td>
</tr>
<tr>
<td>The dog itself</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level</td>
<td>Goals and Objectives Rhetoric</td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Community Mental Health</td>
<td>Schools</td>
</tr>
<tr>
<td>Source of authority</td>
<td>Improve the health, education, and welfare of the U.S. people</td>
<td>Have a better school system in the country</td>
</tr>
<tr>
<td>Those-in-Charge domain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Political appointees</td>
<td>Provide better health for all</td>
<td>Provide a sound, basic education; develop self-actualizing individuals</td>
</tr>
<tr>
<td></td>
<td>Treat alcohol, drug, and mental-health problems</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Improve the mental health of the U.S. people</td>
<td></td>
</tr>
<tr>
<td>Bureaucratic leadership</td>
<td>Make integrated community mental health available to all Americans</td>
<td>Keep all of the schools running in my area, improve test scores</td>
</tr>
<tr>
<td>Working bureaucracy</td>
<td>Meet the mental-health needs of the community</td>
<td>Operate the Burning-Book School</td>
</tr>
<tr>
<td>Direct Intervention domain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First-line supervisor</td>
<td>Keep drug-maintenance patients from drifting off medication</td>
<td>Motivate teachers to balance theory and computation</td>
</tr>
<tr>
<td>Direct intervenor</td>
<td>Maintain Herman on lithium carbonate</td>
<td>Teach Arthur prime factors and how to produce them</td>
</tr>
</tbody>
</table>
implies, the most concretely referenced set of descriptions is usually generated by those in the direct-intervention domain. Even there, however, you will do well to apply some critical judgment—particularly where jargon abounds.

To illustrate, when the conservative tide swept in during 1980, many social-service programs—state and local as well as federal—came under intense scrutiny. Virginia was among the states that took a hard look at its community mental-health program. Those in charge of the program issued some stringent directives to the field offices, among them a directive to have the clients specify what their particular problems and treatment objectives were. Those in charge explained that this would ensure jargon-free responses. However, one of the clients responded that his problem was a large mouse that was living in his stomach; treatment objective: to get rid of the mouse. Presumably, some computer in Richmond has coded and processed the response.

Two examples of semantic shifts are shown in table 5-2. These examples are typical of the rhetoric found in governmental organizations. There are at least three reasons for the language variations that occur among the levels of organization. First, the very purpose of organization in many cases is to incorporate diverse functions, so that more-generalized descriptions are needed at the higher levels. Second, at the higher organizational levels, descriptions of activities are translated from specific detail into the policies and reasons behind them, regardless of differentiation of function. Finally, as people are organizationally and physically farther from the direct activity, a simple semantic drift away from tightly referenced descriptions of activity, affect, and effect occurs.

It is often difficult for the evaluator to determine which kind of semantic change is being dealt with, and especially whether the language is related in any way to the program being evaluated. Considerations must include how the different descriptions affect the determination of what is being dealt with, what questions are to be answered, and how the questions are related to actual activities in the direct-intervention sector. A rough functional model of the direct intervention, constructed early by the evaluator on-site, can often serve as a useful reference in sorting out the rhetoric of those in charge.

The evaluator must deal with each group in the those-in-charge and in the direct-intervention domains in the group’s own language and must carefully collect descriptions, problems, and questions that can be translated into some common framework for use in evaluability assessment and design.

As noted earlier in the chapter, our experience indicates that the best way to make sense of all this is to develop, early, the equivalent-to-reality functional models of the direct-intervention and those-in-charge domains and the family of testable models developed in the those-in-charge domain. These models can be used for describing the direct-intervention domain; the continual elaboration, in the those-in-charge domain, of each group’s rhetorical model of the same actual activities;
and helping to construct models in each unit’s own terms of the activities that they carry out in attempts to facilitate, assist, or enable the direct interventions.

The development and use of extended detail in the equivalent-to-reality functional models provides several benefits:

1. The evaluator will ultimately have to design measurements and discuss attribution of cause and effect to the satisfaction of those in charge. Because the eventual measurements for these expectations must be made in the direct-intervention domain, a defined frame of reference makes it easier to arrive at satisfactory measurements and unambiguous discussions.

2. A functional model, incorporating the evaluator’s best perception of reality and the mechanisms through which that reality can (or might) actually be altered, is the best reference point for sorting out to what extent everyone else is addressing reality in their rhetorical models and to what extent the linkages that they believe they use in affecting the process actually exist and may work.

3. The evaluation design must be heavily conditioned by the information that can be obtained in the direct-intervention domain and by the expectations that those in charge hold, if execution of the design is to lead to purposeful behavior in the organization. The linkages for making changes in the direct-intervention process usually lead through the those-in-charge domain to the intervention. Therefore, it is beneficial to map out the real cause-and-effect possibilities in the those-in-charge domain and to relate them to the functional model of direct-intervention activities.

4. The evaluation designer is trying to match up sets of questions about expectations with possible sets of measurements and comparisons in order to produce an evaluation design that answers the questions. The functional models developed provide the references for this. When a design is completed, of course, the results will have to be translated back into the particular language and terms used at each particular level, communicating what has been done. The very process of examining the those-in-charge domain and explaining the work as it progresses also provides a unique opportunity for the evaluator to condition those in charge to use the information that the evaluation produces.

The implication-loaded language, the overabundance of feedback loops, the multiple levels of organization and semantics, and the (sometimes sharply) different models held of what is actually occurring at the working level are the normal facets of any large organization that evaluation designers (as well as everyone else) must deal with routinely in everyday work. Do not be nonplused at encountering them. It is not much use to simply state, “Be sure to determine that the important concerns held by management are selected for evaluation,” and to stop at that point. To improve purposeful behavior in government, consideration
of the factors described in this chapter is always necessary. Ignoring them inescapably produces a futile evaluation.

Note

Chapter 5 was concerned with those in charge and how the evaluators begin to cope with them. This chapter briefly discusses how those in charge cope with evaluators.

Early in the evaluation, those in charge form perceptions of what the evaluators are doing. Those perceptions have important impacts on the success of the evaluation. Specifically, the early perceptions influence what those in charge are willing to tell the evaluators and, ultimately, whether the evaluation results are even read, let alone used in support of purposeful management behavior. Thus, it behooves the evaluator to consider carefully the factors that shape the perceptions of those in charge.

Why Is Acceptability a Problem?

The major influence on the opinions formed by those in charge is the fact that evaluation, one way or another, measures performance. Its end product is a rating, an assessment, a judgment. Very few people like to be judged, which is one of the hard facts with which the evaluators must deal. Another fact is that if people must be judged, they will do everything in their power (which may be considerable) to see that measures are selected to show them in a favorable light. Each level of the organization will lobby for the use of its own preferred measure rather than someone else’s. Thus, the mere suggestion that an evaluation is in the works can stir up a multitude of animosities among those in charge as they jockey for a favorable evaluation position. In an article noting this behavior, Aaron Wildavsky remarked, “I started out thinking it was bad for organizations not to evaluate, and ended by wondering why they ever do it.”

The evaluator can often allay the fear of judgment by being aware, in advance, of what kinds of measures and comparisons are likely to be accepted under different conditions. The evaluator must gain some idea of what those in charge will accept as performance measures. Such measures can then be considered in the evaluation design with some confidence that, if acceptable measures prove feasible to obtain, the results may be valuable. At the very least, the evaluator can work—during the design effort—to begin to develop a common context for discussions among those in charge about performance. The development of common contexts will pave the way for discussions of results once the evaluation is complete and for the use of the findings to produce program improvement. The trick is to solve Wildavsky’s riddle,
“How might analytic integrity be combined with political efficacy?”

Once committed to an approach based on improving purposeful behavior, the evaluator begins to define carefully what is being done, to establish measures to be taken, and to perform analyses and comparisons. This is all done with one eye to producing and feeding back the resulting information to those in charge in a way that will influence subsequent actions to improve the measurable performance and with the other eye on the technical soundness of the approach. Those in charge, however, are well aware that evaluation results are not for their perusal alone. Whatever the organizational level, those in charge are usually not free simply to examine the evaluation results and act on them. Someone above them will also read the results, and a negative evaluation can lead to program dissolution rather than program improvement. Thus, the people most affected by the evaluation may be enormously sensitive to how the evaluators intend to keep score.

If those in charge form the early perception that the evaluators are going to select simplistic measures and comparisons, further cooperation may well be foreclosed (except in those cases where simplistic measures are likely to show simplisticly that those in charge are doing an excellent job). In order to build up and to maintain the credibility of the work, the evaluators must draw out and incorporate the diverse expectations and goals of those in charge into the design. Four common mistakes in determining the acceptability of assessment measures include viewing the organization as:

- Having a single view of what the service, operation, or direct intervention really is;
- Having a single perception of cause and effect;
- Being represented by a single group or person or as having a single set of goals;
- Accepting (or having to accept) a single comparison standard or type of standard.

In order to get an idea of how rich the possibilities for the selection of measures are, even for simple examples of process activities, examine the sample worksheet in figure 6-1. Along the bottom have been shown cost and existence of intervention, process measures, immediate outcomes of process, and impact measures. For any particular intervention or process, one could begin laying out characteristics to be measured. Along the left side have been shown some typical comparisons that might be made once the measures are taken. They range over fixed standards, what was planned, what was done last year, relative comparisons of the same measures with other similar activities, peer-group assessments of
results, and professional standards of behavior. A great many more measures and comparisons can obviously be generated—for example, percentage of recruits passing the test, resources spent per recruit, or comparisons with other battalions. Figure 6-1 suffices, however, to indicate the potential (and necessity) for complexity.

Clearly, not all boxes would be considered in every case, and some would only be used in a few cases, but the possibilities inherent in even the simplest evaluation are quite large. As an example, consider once again a home heating system as a model of a direct intervention. For the sake of this example, we will evaluate the furnace from its date of purchase and regard the levels of those in charge as including not only the family but also the people who were responsible for installing the furnace, servicing it, and insuring its warranty. Figure 6-2 shows some of the possible entries for the boxes in figure 6-1.

Obviously, the furnace installer would resist being evaluated in terms of the number of colds the family caught in a given year. Also, the warrantor is unlikely to look with favor on a measure that depends on the comfort of a few sedentary bridge players. This is particularly true if the people involved feel that continued business depends on a favorable evaluation.

<table>
<thead>
<tr>
<th>Some Possible Measures and Comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional Standards of Behavior</td>
</tr>
<tr>
<td>Peer-Group Assessments of Results</td>
</tr>
<tr>
<td>Measures of Other Similar Activities</td>
</tr>
<tr>
<td>What Was Done Last Year?</td>
</tr>
<tr>
<td>What Was Planned?</td>
</tr>
<tr>
<td>Fixed Standards</td>
</tr>
</tbody>
</table>

Figure 6-1. Examples of Some Possible Measures and Comparisons
One other subtler point should be noted. While the furnace installers might want to be evaluated on a fixed standard—namely, the physical presence of an operating furnace in the basement—the installers may not believe that the standard will be accepted by potential customers (after all, everyone has an operating furnace). Therefore, the installers may not believe that the standard will be accepted by potential customers (after
all, everyone has an operating furnace). Therefore, the installers may prefer a somewhat chancier measure—for example, the price of their furnaces compared to the price of others on the block or whether most people think it runs quietly.

Transferring this concept back to the evaluation of a complex social program, three things become clear:

1. Each level of the those-in-charge domain may prefer measures over which they feel that they have some control and comparisons in which they have a chance of shining.
2. Each level of the those-in-charge domain will prefer measures and comparisons that will be accepted by so-called significant others—people on whom those in charge depend for money, guidelines, or information.
3. An acceptable range of possible measures is unlikely to be generated simply by surveying a few decision makers and asking them to list the questions they want answered.

**Factors Influencing Acceptability**

Some groups among those in charge will strongly reject some measures while they accept others. Different groups may reverse the preferences. Probably only the staff researchers will accept everything.

It is possible to isolate some of the key factors that determine what measures and comparisons will be accepted by those in charge. Two major factors that influence acceptability are (1) whether those in charge believe that they can predict with relative certainty what significant others will accept as evaluative evidence and (2) whether those in charge believe that a predictable cause-effect relationship exists between their actions, the direct intervention, and the desired outcome—that is, are they sure that they can make their program work?

Three possible states might pertain to each of the factors. For instance, those in charge might believe that the cause-effect relationships are

Well known,
Uncertain but plausible,
Implausible or hopelessly confusing.

Those in charge might believe that significant others:

Will or will not accept given measures and comparisons (certain belief),

Might accept given measures and comparisons (ambiguous belief),
Will behave unpredictably (the probability of acceptance is completely unknown).

The relationship between the beliefs about cause and effect and the beliefs about significant others' acceptance of measures is likely to have considerable effect on what measures and comparisons those in charge will accept. The possible relationships can be more easily explored by referring to the matrix shown in figure 6-3.

At (A) in figure 6-3, both the cause-effect relationships and the evaluative evidence acceptable to significant others are believed to be well known. The congruence of those two variables usually calls for a computational strategy for evaluation. This combination of variables often occurs when the most basic questions are being asked. For instance, a job training program might be teaching people to weld. A certified vocational-education teacher has been employed. The class has been prescreened so that it consists of people sufficiently intelligent and coordinated to learn. Those in charge believe that a given number of instruction hours will cause a given proportion of the class to become proficient welders. Those in charge also believe that significant others will accept a proficiency test in welding as evidence of learning. Ergo, that is a computational strategy—administer the test and count the number of students who pass.
At (E) in figure 6-3, the cause-effect relationships are uncertain, and those in charge are not sure what significant others will accept as performance measures. This situation might pertain, for instance, if the question being asked was not whether people are learning to weld but whether the job training program is reducing unemployment in Toledo. Here, an iterative strategy for both management and evaluation is probably in order—that is, the evaluators must simply engage in series of discussions with those in charge, proffering and refining suggestions for acceptable measures at any stage of the program until (one hopes) inspiration strikes and ultimate agreement is reached. Similarly, it may be necessary to build up the evaluation, starting with simple monitoring, moving later to immediate outcomes, and only finally developing potential impact linkages as more and more is known. The problem may be too complex and too little understood for a complete measurement model to be immediately laid out and agreed to in advance.

At (D) in figure 6-3, cause-effect is believed to be understood, but there is doubt about what performance measures significant others will accept. Public-school principals often hold this combination of beliefs. They believe that they know how to select and hire good teachers. They are not sure that the school board will accept reading proficiency as a performance measure (social integration, for instance, may be a stated goal). In this situation, a comparison strategy is recommended—that is, given the fact that the principals often do not know what significant others will expect of their school (or accept as a measure), they will usually prefer to be compared with other principals with schools like theirs rather than with an unknown standard. They will rely upon their understanding of their work to remain competitive with their peers.

At (B) in figure 6-3, the situation is reversed. Here, those in charge know what is expected, but they are not sure that their intervention will cause the desired outcomes, although they believe it may. An example of this might be a program that sponsors halfway houses for former drug addicts. The desired outcome is quite clear: The former addicts should spend the rest of their lives free of drugs. The people who sponsor the halfway houses, however, are usually not sure that their treatment causes the desired outcome. In this situation, those in charge tend toward one of two gambling strategies. Either they will want to be measured early in the process by a proxy measure (for example, number of people accepted for treatment), gambling that they can convince significant others to accept the measure, or they will want to be measured long after the process is complete (for example, percentage of former addicts who are drug free after five years), gambling that either the intervention really does cause the desired outcome or that some unknown intervening variables will help the results (or at least make them ambiguous) or that they will be
gone before the results are in. In any case, they will have avoided testing the doubtful cause-effect relationship immediately.

We have extended our matrix beyond that used by Thompson to include situations in which the goals acceptable by significant others are actually believed to be unpredictable and changeable. Over the last few years, a combination of changing views in Congress, short-term political appointees in the executive branch, and rapidly changing or conflicting views of proper goals by citizens have led to changes in what is acceptable in many areas and at many levels. The situation may spring from simple conflict as in day care, from sharp program change as in manpower, and apparent philosophical changes as in welfare, education, corrections, and the courts. It is often easier for the leadership to change their views on what is acceptable than for direct intervenors to change what they are doing. Thus, policy switches create a certain suspicion of evaluation on higher-level expectations and goals, especially if the evaluation effort appears to be a serious one.

While we do not discuss all of the associated boxes [(I) on figure 6-3 is dealt with later], the evaluation designer should be prepared to encounter some cases in which those in charge believe (sometimes correctly) that, by the time a lengthy evaluation is completed (1-3 years), the measures selected will be unacceptable because of new congressional thrusts, new leadership, or governmental idiosyncracies.

In other cases, legislative bodies (whether Congress or the local school board) continue to authorize programs that are statements of intent where effective methods of intervention may not be widely known—for example, raise the mental health of the country, improve education, reduce crime, end inflation and unemployment, cause economic development among the poor, or eliminate dependency on welfare.

These programs are authorized when a consensus is reached that something must be done but when the method for doing it is not clear. Such programs often call for much innovation in developing an approach. Where this occurs, many people associated with the program will believe that they do not know a method of producing the result or that the prescribed method is doubtful. Occasionally, the extreme case such as at (I) in figure 6-3 will be encountered where many people believe that the acceptable measures will change and that the efficacy of the methods of intervention are doubtful. Some present examples might be the information programs to induce voluntary reduction of energy consumption, to reduce crime, or to provide certain types of health care.

In many cases, the evaluation designer should not be surprised to find many middle-level bureaucrats who tenaciously resist most program-oriented measures. Like any sane person, the bureaucrat may attempt to push the evaluation to mundane things that may visibly happen:

What Is Acceptable to Those in Charge

We met 40 of our 60 process objectives.
We established 22 projects, and all paper work complies with agency and OMB requirements.

We moved our money on time.

We have distributed descriptions of exemplary situations.

We are building a cadre of skilled workers (researchers, trainers) in this area.

In some cases, the area simply may not be suitable for evaluations (or for management). In others, what may be required is a formative evaluation system that is developed as the intervention is developed—that is, a system that begins with monitoring of existence as things begin to exist, monitoring of process as processes develop, evaluation of outcomes as they occur, and potential impact measures as impact models can be developed. In such cases, the evaluator must be extremely sensitive to the acceptance of those in charge and to the nature of the intervention so that measures can be constructed that are appropriate at each stage of development and that can be used in further management and development of manageability. Manageability and evaluability are intimately related.

This is quite different from the classic experimental design. What is under discussion is evaluation as a part of a process for moving a governmental program slowly toward (A) in figure 6-3 (where certainty about effects coincides with certainty about acceptable measures) from some more-ambiguous or possibly even chaotic location such as (I). Note that such processes must often start with simple monitoring and move on to more-intensive evaluation. Interestingly, if the program ever reaches (A) of figure 6-3 (where both accepted measures and knowledge of cause and effect are both believed to be well known), simple monitoring to see that it all comes true may be enough once again.

This discussion illustrates that where high degrees of certainty pertain, those in charge may be willing (even eager) to settle on concrete performance measures and comparison standards. Cause-effect assumptions can be tested (for example, individual tutoring raises achievement scores) through effectiveness measures. Relative efficiency measures can be applied (for example, can the same results be attained by using paraprofessionals rather than highly paid professionals?). As uncertainties about cause-effect creep in, those in charge may turn to comparisons (How are we doing compared to last year? or How are we doing compared to Illinois?). If uncertainties about goal expectations are coupled with confidence in their ability to cause specific results, those in charge may push for comparisons of planned performance versus actual or comparisons with peers on the same measures.
As uncertainty meets uncertainty, those in charge may evince more interest in peer-group assessment (They are the only ones who are capable of understanding problems like mine) and professional standards. An interesting variation is the substitution of organizational standards for professional standards, usually occurring where professional standards are nonexistent or cannot be met. This is common in large governmental organizations. Organizational rules such as a requirement for all personnel to be on-site from 8:30 A.M. to 5:00 P.M. are created. There may or may not be a clear idea of what compliance with the rule is supposed to accomplish, and there is almost certainly no validation of the assumption that compliance does accomplish it—whatever it is. Nonetheless, the organizational standard, once created, takes on the patina of a valid measure, and organizations point with pride at their success in meeting the standard (82 percent of the faculty hold Ph.D. degrees) and with less enthusiasm for outcome or impact measures (The furnace exceeds American Institute of Architects (MA) standards! What do you mean by “Keeps the house comfortable, anyway?”).

Analytic Integrity and Political Efficacy

Earlier, Wildavsky’s riddle was quoted, “How might analytic integrity be combined with political efficacy?” The ensuing discussion has been taken up with questions of political efficacy—that is, the identification of measures likely to be acceptable to those in charge. The emphasis placed on politics should not be interpreted as an abandonment of integrity. Indeed, the rest of this book is concerned, one way or another, with the construction of honest and useful evaluations. The purpose of this chapter is to introduce the evaluator to some of the organizational mouse traps that may await. The prize mouse is the one who avoids the traps, selecting multiple measures and comparisons that will be acceptable to those in charge, lead to reasonable appraisals of program performance, and cause adjustments based on those appraisals. Forewarned is forearmed.

Notes

2. It may be helpful to review some of the material addressing the importance of common contexts to communication between persons in large organizations. Chester Barnard, in The Functions of the Executive (Cambridge, Mass.: Harvard University Press, 1938), presented what remains one of the best treatments of the subject. See particularly page 10 of that work.

What Is Acceptable to Those in Charge

4. The following discussion is an adaptation and extension of a concept


Some Clues on Exploring
In the Those-in-Charge
Domain

In chapter 1, we noted that evaluation is used to support purposeful management behavior and that an important activity carried out by evaluators is the translation of raw data into a form understandable by busy senior people. Two things follow from those statements. First, the evaluator must determine who are potentially successful users of information—that is, which people have the motivation, understanding, ability, and authority to apply the results of evaluation in ways that will (or might) improve direct-intervention activities. Second, the evaluators must learn the languages of those in charge and establish a common context in which the evaluation results can be reported and mutually understood.

Given normal limitations on resources, the evaluator will not be able to interview everyone in the organization. The questions that naturally arise are:

How do the evaluators know which people to talk to? What do the evaluators do once the significant people are identified? Generally speaking, an efficient way to identify significant actors is to trace an important flow (for example, information or money) through the organization and to interview people in the path of that flow. This will provide a first cut.

In the initial interviews, the evaluators will begin to develop a familiarity with the languages of those in charge and—just as important—will be able to fine tune the identification of significant actors. It is not uncommon to find that some people who seem to be important—who stand squarely in the path of an essential flow—in fact simply pass things along from one decision maker to another without adding (or subtracting) anything substantive of their own. As these people are recognized, the time spent on them can be reduced and added to other areas. Few should be chosen for detailed examination, although many may be called.

The inhabitants of the those-in-charge domain are usually organized into several layers and offices, as pointed out in chapter 5. There will be differences in what information they want, will accept, or can use. This chapter presents some clues that may help in identifying the appropriate people and extracting their requirements.

It is important to find those few members of the those-in-charge domain who, by authority or strategic position, may actually affect (or have some link to) the direct intervention. An evaluation performed for an office from the those-in-charge domain that has no connection to the
direct intervention offers less chance of causing an effect or establishing a completed loop—that is, the resulting evaluation is less likely to be used. (Remember that a purposeful-behavior loop is intended to extend from performance through measurement to management and back through management actions to a new level of performance.)

After having found the actors who might affect the direct intervention, it is still necessary to learn their particular languages of description and decision in the area of concern. The ability to speak to actors in their own language goes a long way toward dispelling the natural distrust inherent in dealing with people from an alien domain. The very process of interacting with the members of the organization on their own terms may lead to a favorable reception of the evaluation results when they are produced. At the very least, they will be read by some of the significant people; at best, they will be used intelligently.

Unfortunately, many texts and courses in public administration—particularly those courses in statistics required of incipient evaluators—stress the need for managers to learn the language of quantitative analysis. Not unnaturally, the young evaluators feel that the senior people either are, or should be, technically literate and that, if the evaluation report is misunderstood or ignored, it is the fault of those in charge. Since most senior people have probably never taken a course in public administration, they remain invincibly ignorant of their linguistic obligations to the evaluation domain and steadfast in their refusal to take actions on the basis of reports they do not understand and know they do not understand.

The more-technical part of an evaluation will require the selection of measures and comparison standards, specifications of acceptable accuracies, and estimates of the trade-offs between the costs of obtaining information and the values of the potential uses of that information. All of these questions affect design, and an important fact is that the answers are usually—at the beginning of such an investigation—distributed loosely among the people in the those-in-charge domain. Many people can be found who will answer such questions definitively, but their answers may each be quite different. Who should be listened to? Whose answers should be developed in more detail?

Having belabored the importance of identifying, understanding, and learning the language of a few, significant members of the those-in-charge domain, it would seem that what should follow is a reliable, valid method that without fail establishes who and where such people area. Our experience indicates, unfortunately, that this is not possible. Each new case seems to produce a different pattern; each agency seems to have a different de facto trail of influence.

As we mentioned in Chapter 5, traveling to various parts of the those-in-charge domain of the same governmental organization often resembles island hopping in the Pacific. Each new insular group encountered displays different understandings of its own situation and of
the world, different subcultures, different customs, and different
languages. The evaluators must master these various tribal customs
before discussions of each unit’s program, the unit’s part in it, and the
unit’s information needs can be carried out meaningfully. Without such
mastery, the evaluators can be misled and can err badly. Because the dif-
ficulties of island hopping are so pervasive, the remainder of this chapter
is devoted to providing clues to island hopping without drowning. These
are, however, only clues, not sea charts. Our state of the art has not yet
reached the point at which specific solutions can be prescribed.

The Treasure Hunt

There are seven clues for the treasure hunt. The first two involve
finding the right people to talk with. In order to do this, the evaluators
must range through the organization. The latter five clues pertain to what
evaluators do after the appropriate people are identified, when the
evaluators are dealing with specific members of the those-in-charge
domain. The seven clues involve

Information flow;
Funds flow;
Scope of discussions;
Detailed content and semantics of discussion;
Decisions, resources, and orientation;
Patterns of time use;
Separation of effectiveness and efficiency discussions.

Each clue is discussed in turn in the following sections.

Information Flow

Searching out the information flows around the topic of concern
and tracing how these flows presently take place within the organization
is often a good place to start. Combined with the flow of funds discussed
in the next section, the information flow often gives one an overview of
where to go later with more detailed questions. Authority structures, per
se, have been omitted from this list of clues since a little initial reading
and an organization chart usually gives the evaluator the purported
authority structure. Because the actual (or operational) structure is often
Based on quite different informal lines, it is seldom fruitful to rely on the formal organization charts and vested authorities. Success is more likely to come from tracing the information (and funds) flows and inferring the authority from those. This is especially true where content knowledge is involved. An information flow is the movement of pertinent facts and concepts from one location to another. The flows can go downward—as in the cases of program guidelines, policies, or management by objective (MBO) system plans. Information on process or performance (if the information exists at all) is expected to flow upward. At some levels, important horizontal flows exist. The approach here is to loosely locate those myriad flows, fasten onto particular artifacts of each flow (documents, reports, briefings, and even raw data), and ask at each way station questions such as:

- Where did this information come from?
- Why did it come here?
- Is part of this information created here?
- Was this information transformed in some way here? How?
- Where does it go next?
- Why does it go there?

As answers to the questions accumulate, the evaluators can begin to rough out a picture of the actual information flows against the formal organization chart.

**Funds (or Funds-Authorization) Flow**

The funds usually originate at the place at which the appropriations take place (for example, Congress or a school board) and eventually arrive at the level of the direct intervention. Unlike the information flow, which may either swell or contract at each way station, the money invariably diminishes as it flows downward through the agency, reduced by the amount of overhead activity being carried out. Also as it travels down, the terms of its ultimate use are often constrained and structured by guidelines and regulations that are fed into the information flow in response to authorizing legislation or to good-practice beliefs of those in charge. Making the attempt to trace the funds flow through a governmental agency, while often tedious, is almost always rewarding. The location and nature of commitment, proposals, and approval and the various release points and authorizing persons—when
combined with the information flows collected—often contain the best indications of the de facto authority structure of the organization.

In the funds-flow patterns governmental organizations differ sharply from industrial organizations. The analyst who is attempting to transfer techniques from industry (Let’s apply good business management here and clean up this mess) must be aware of a simple, yet fundamental, difference between funds flow in much of industry and funds flow in many governmental organizations. In much of straight commercial industry, money flows up through the organization—that is, there are production or service operations (the equivalent of a direct intervention) that are organizationally at the bottom of the company but that bring in the money for the organization. In many cases this money can be identified with the unit producing it. Whatever other factors private-sector decision makers would like to consider, they cannot completely avoid, over any extended period of time, the question of whether or not a division or section is profitable, or at least self-sustaining. If enough money to at least support that part of the organization ceases to be produced, there is usually a limit to the amount of time over which someone will send money into the unit to keep it running. Since the funds are supposed to be generated by the activity, profit (or at least self-sufficiency) becomes an enforced performance measure that is an unavoidable factor in many decision considerations. [Contrary to common mythology, however, profit (or break-even) is far from being the only performance measure; it is simply an omnipresent self-enforcing indication that one condition of success has been accomplished satisfactorily.]

In governmental organizations (excepting fee-for-service arrangements) the flow-of-funds situation is the opposite. The market for producing the money is found in a different feedback loop. This loop contains the appropriations source and the people who influence it. At the federal level, an agency may only have to sell a subcommittee or two in the Congress, its own lobbies and constituencies, and the media as long as the general public is not aroused or vitally interested. (Note that the selling is done, at least in part, by people from the those-in-charge domain. The direct intervenors do not deal with Congress unless their testimony is requested by the committee, although a few carefully selected practitioners may be asked show and tell in order to dramatize a point or to lend an air of reality to the proceedings.) The money appropriated then flows down through the organization and is allocated to different specific uses. In chapter 8, some of the effects of this downward flow on incentives within the organization are discussed in detail. Suffice it to say here that this process must be understood in order to trace the flow of funds and to understand the
often puzzling content of the information flow that accompanies the funds. Many of the regulations, guidelines, and procedural instructions are attempts to satisfy a subcommittee (often one bent on accountability) rather than to improve the direct intervention or the agency’s own administrative practices.

**Scope of Discussion**

When those in charge discuss their program, careful notice should be taken of the range, or scope, of the issues they discuss, the actions they consider to be available, the problems they see as important, and the performance information they already know about or are interested in. The range of the program as carried in their minds, the possibilities of action as perceived by them, and the range of activities in which they see themselves as being involved help to delimit their particular island as they see it. Combined with other intelligence, such information contains essential clues both to the further interaction necessary to enable later use of the evaluation results and to parameters of design. It is often helpful to sit in as an observer at some policy meetings considered especially important by members of the unit.

**Detailed Content and Semantics of Discussion**

The scope of the discussion defines the outer boundaries of the program as each group sees them, such as the issues involved and the kinds of decisions made. Within those boundaries, the detailed content and semantics of discussions help the evaluators to understand the organizational integration (or isolation) of the unit (compared to other units and to the direct intervention) and to comprehend the often ambiguous language accumulated in successive interviews. It must be kept in mind that at each group and organizational level, even common English words may have a tendency to take on specialized local meanings, and the evaluators may be, in a real sense, learning a local language. For example, consider what citizen participation might mean to different units in the those-in-charge domain. To some, it is written communication from a group with a special interest in particular regulations; to others, it is the complaints or plaudits of unaffiliated citizens responding to statements of policy or posture; still other units regard citizen participation as real only when local citizens become involved at the direct-intervention level by serving as volunteers, on advisory committees, or as paid employees of the project in question; some units (albeit not many) see citizen participation as properly coming during the evaluation phase of a given activity when the opinions of local citizens are solicited.1

Comparisons of content and semantics made across the organiza-
tion can help to determine how much common context is available (or needs to be created) and to what extent useful information has to be tailored or interpreted for each group. Early in the evaluation design, the evaluators will also have to compare the various models of reality held by those in charge with the evaluators’ own equivalent-to-reality models in order to begin to reconcile questions being asked by those in charge with answers that can be obtained by measuring observable activities. Actual samples of semantic usage often become invaluable in this process and should be preserved in verbatim interview notes.

Decisions, Resources, and Orientation

Each group in the organization has a set of beliefs (sometimes but not always accurate) about what decisions it makes and what resources it controls. Also,

- each group is oriented (direction facing) either downward toward the direct intervention, upward toward the source of authority, or in a Januslike combination of both. These beliefs and orientations are of interest since they shape information needs and uses. The semantics and scope of discussions are often helpful in ferreting out this information.
- The discussion of citizen participation, for instance, differentiated several possible meanings of the term. Accepting quick answers or survey-type information without exploring what the words mean in each case may be very misleading. The meanings chosen for particular words by any given group may provide unmistakable cues about its direction-facing and some indication of its beliefs about its own decision powers. Again, sitting in on some meetings may be a good addition to straight interviewing.

Patterns of Time Use

In all of these collections, a time will come for synthesis and comparative analysis in order to develop an overall picture of both the rhetoric of those in charge and of their observable or believed activities. While much of this can be done by relating the various parts of the organization to each other, it is often helpful to develop some reality tests to use judiciously against material obtained through straight interviews or observation. In the introduction to this part on those in charge, we described the several types of models that are useful in conducting this part of the work. These models included the family of testable models (logic, functional, and measurement) that are drawn solely from the descriptions of those in charge, and the equivalent functional models of the direct-intervention and the those-in-charge domains that are based on the evaluator’s observations. Over time, these models will have to be compared and the discrepancies resolved.
In resolving the discrepancies, it helps to locate some touchstones of reality to serve as checkpoints. Unobtrusive measures such as records of past actions over time, past decisions, and past advocacy patterns or documents are useful. One of the best checks, however, is the actual pattern of time use by the people involved. The methods to be used depend upon the time available and the experience of the investigators. Bear in mind, though, that simple interview questions are almost never sufficient. Secretaries’ desk calendars or schedules can often be important. Analysis of the activity of key actors over a few weeks will often reveal what duties actually claim their time as opposed to what they say claims their time. Observation is sometimes possible, although it is often tedious and expensive. Procedures such as structured observations are often useful. Logs, showing time expenditure over some period, kept by those in charge are also helpful, although logs tend to be more accurate if kept by a secretary. The idea is to develop a few key time-allocation patterns to see if they fit all of the interview material collected.

Effectiveness and Efficiency

Material developed in discussions of programs often contains two different kinds of questions and approaches (effectiveness and efficiency), and it is useful to distinguish between them. Drucker, among others, gives examples of efficiency and effectiveness. He understands effectiveness to mean that the option chosen is the one that is going to be most right for your market and takes efficiency to mean doing whatever is being done in the most careful, least costly way. To further distinguish them, he points out that no amount of efficiency would have kept buggy-whip manufacturers from going out of business. For the buggy-whip manufacturer, new measures of effectiveness were needed in order to serve a new market. That is a quite different thing from improving your efficiency at making buggy whips. Only in special cases do changes in simple efficiency make or break an industrial company. The way many companies keep going is by adapting their product lines to the exigencies of the market. Those concept changes are so much more important than operating efficiently that careful limits to the amount of time spent on improving operating efficiency versus effectiveness must always be considered.

Why, then, do so many government agencies concentrate on efficiency studies and measures? On the one hand, they do so partly because efficiency studies are usually easier to do, much easier to justify, and quite safe because some results can almost always be obtained. On the other hand, effectiveness studies are much harder and have extreme chances for failure or success. In government, where appropriations support for a program is often independent of the effectiveness of a program over a period of time, those in charge may only be interested in
how a program’s efficiency can be improved. Indeed, this may be the only information they are able to act on.

In interviewing the significant actors, try to distinguish between those whose interest and clout are limited to efficiency considerations (How well do I do what I do?) and those, if any, who can influence effectiveness (Do the things I do have the outcomes and impacts desired? Are there better approaches?). The measurement and comparison techniques needed may be quite different for each.

Remember that, given normal limitations on resources, the evaluation designer will not be able to draw out the views of everyone in the organization. Locating those actors most (and by contrast, least) likely to make use of the subsequent information vastly increases the chances that the evaluation results will be applied to program improvement.

Notes

1. An example of citizen evaluation is the study reported by Daniel Katz, Barbara Gutek, Robert Kahn, and Eugenio Barton, in Bureaucratic Encounters(Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, 1975).


3. A good description of structured observation, including a comparison with other similar techniques, can be found in Henry Mintzberg, The Nature of Managerial Work (New York: Harper & Row, 1973).

In designing, implementing, and reporting the results of an evaluation effort, two things become evident. On the one hand, deciding upon standard information, getting it collected and fed back, establishing expectations and comparison standards, and getting those involved to make the comparisons between performance and standards are difficult, but manageable, tasks. On the other hand, designing such efforts so that action will be taken by those who make the comparisons is infinitely harder. Indeed, a major topic of discussion among evaluators and evaluates alike is why and whose fault it is that evaluation results are seldom used to support purposeful management behavior.¹

Throughout this book, we have dealt with the subject of utilization from several perspectives. In particular, chapters 5, 6, and 7 discussed what to look for when dealing with those in charge so that usable, used evaluations can be designed. It should be understood, however, that even the best designed evaluations will probably not have impact right away. Except in rare cases, reinforcement over time will be necessary if evaluation results are to be used in support of purposeful management behavior. In order to grasp why this is true, the evaluator should not only know what those in charge are like but also how they got that way. This chapter thus describes the so-called incentive framework of a governmental operation.

In chapter 2 we discussed purposeful behavior in the sense that Herbert Simon uses the term; that is, behavior that is intended to gain an objective and that is continually corrected as errors are detected.² During the adjustment process, certain patterns of behavior are learned (or internalized) to the extent that they become automatic responses to given classes of situations. Sometimes the learned responses are not always appropriate. If a child learned to walk, for instance, in a house that had bumps in the floor in front of every door, the child would probably learn to give a little hop every time he came to a doorway. This behavior, however inappropriate, might stay with the child throughout life since one can hop through a doorway without much sacrifice in efficiency.

Most people have some idiosyncratic behavior patterns, both appropriate and inappropriate. They bring these behavior patterns with them to any organization they join. This learned behavior forms part of the store of information that people use when called on to respond to
particular situations. If a person is particularly influential (for reasons such as status, force of personality, or ability), the stamp of that person’s idiosyncratic behavior may virtually pervade the organization.

Superimposed on the idiosyncratic stored knowledge of its members are the behavior patterns encouraged by the organization itself—that is, members of the organization, if they belong to it for a period of time, learn that, for them, some kinds of behavior are rewarded with success in forms such as applause, money, or promotions. Other kinds of behavior cause a lot of trouble in forms such as irate clients, withdrawal of support, or a succession of losing battles. The members’ perceptions of the behavior that leads to success or failure cause them to adjust their stores of knowledge about how to achieve desired purposes. Also, information gained through successive learning experiences within the organization seem to have much more influence on what is done by individuals than circulars or guidelines about goals and objectives.

Thus, when individuals are presented with the plan for, or the results of, an evaluation, their responses to it are filtered through their idiosyncratic stored knowledge including the knowledge acquired as part of their organizational experience.

Much of the store of knowledge acquired by an individual as part of the organizational experience is shared by other members of the organization. When this store is shared by many members it becomes the basis of a tide of organizational behavior that is soon seen as characteristic of that organization. Often, the idiosyncratic behavior patterns of individual members of the organization will be only partially submerged by the tide of the distributed organizational response, causing minor—but often interesting—perturbations in the directional stream. Some indications of both the tide of organizational behavior and the idiosyncratic responses of key members are usually uncovered during evaluability assessment (described in Part IV).

**Where Organizational Incentives Come From**

A major market of any organization is the group that must be satisfied if sufficient money is to be generated to keep the organization in existence. This market is a bounding constraint that the organization ignores only at its peril. The market must be satisfied or—over a period of time—it will affect the organization disastrously.

There are then bounding constraints on the types of internalized behavior that ignore the money-generating market, since (over time) such negligence becomes increasingly painful. Those having the most contact with the market are likely to notice adverse changes in market reaction soonest. This is not to say that the organization will necessarily alter its internally stored knowledge and reactions or even quickly change its reinforcement pattern for individuals as a result of this conflict. A wrong-
headed internalization (because it is distributed among many people) may
survive even monumentally adverse market reactions, especially if an
organizationally suitable scapegoat can be found to explain failures. The
one clear point is that, unless operating funds can be generated, the
organization will not be able to continue operating over a long period of
time. If it continually makes responses that are wrong for its money-
generating market, the operating monies will eventually be wiped out.

If the money-generating market and the activities of the organization
are directly related and understood, then the chances of internalizing
survival behavior are good, although far from assured. If the activities of
major portions of the organization and the money-generating market are
sharply separated, with different markets exerting primary influence on
different clusters of individuals, then the distinct clusters will internalize
disparate responses. In other words, one will see an organization at war
with itself. This is a common phenomenon in many federal mission
agencies wherein the so-called service crazies do daily battle with the so-
called political hacks. If the predominant tide of organizational response
is such that the money-generating market is dissatisfied, one will see an
organization with a death wish.

Success for an individual or an organization is an incentive to repeat
a particular behavior pattern whenever a situation arises that resembles
the situation in which the success occurred. Similarly, failure is a
disincentive that persuades people to alter their behavior. The developer
of successful information-feedback devices must consider the reasons
why different sources are perceived (by those who work in government)
as providing the incentives and disincentives—that is, whom is it
important to please?

Earlier, we noted that a fundamental difference exists between most
governmental organizations and private industry. In private industry, the
money that is necessary for the survival of the organization flows upward
from the customers or market. One way or another, the economic survival
of a competitive private-organization depends on the satisfaction of those
customers. This does not mean that customers are the only people that
have to be satisfied, significant members of the organization’s
environment, such as stockholders, bankers, and regulatory agencies, also
require their due. In the final analysis, however, for the organization as a
whole, the customers provide the overridingly important incentive—
namely, survival. (In a sense, the very attempt to build up evaluation
capabilities in governmental organizations is an attempt to develop a
substitute for the self-disciplining feature of industry—that is, paying
your own way as a result of your own direct-intervention activities.)

In a governmental organization, the money mostly flows downward,
from the source of authority. This has some rather odd repercussions on
the people who work for the organization. Some people, particularly
those who are (organizationally) close to the source of authority, quite
correctly perceive that the survival of the organization depends on the
happiness of those providing the money (at the federal level, usually these are the appropriate congressional subcommittees). In addition to being the sources of the ultimate incentive, members of Congress (or their staffs) are the source of the bulk of disincentives. They are the ones who make the phone calls that have to be answered, request the administrator’s presence at hearings, call for investigations by the GAO, say flattering or unflattering things to the press. In short, the stored memory of those parts of the organization closest to the source of authority indicates that the organization should regard Congress as the fount of incentives and the most important market to please.

The fact that the money flows downward often conditions actors in the system to have a large measure of ambiguity in their desire or need for performance information (especially where rhetorical goals may not have been met). We also suspect that it is one of the conditions that cause disasters for those expecting performance and productivity systems that were excellent in industry to work in government.

Many bureaucrats have correctly perceived that definitive, accurate, reliable performance information has not always been necessary to produce the money in the past and that negative evaluations are often deadly to repeated funding. This facet of the governmental environment requires careful examination by the evaluation designer to discover where needs for information actually are felt, what those needs are, and whether anyone is likely to act on the information if it is actually created. A government actor who has no need of performance information in order to obtain the next appropriation of funds may be singularly uninterested in the design of an evaluation unless the evaluation is likely to be done anyhow and may produce negative results or bad publicity.

This theme has a number of interesting variations. One of them was played out by the now-defunct Research Applied to National Needs (RANN) directorate of the National Science Foundation created by Congress to fund applied research projects that would identify areas of national need and apply the results of basic research to the identified areas of national need. The working-bureaucracy level (the program managers) of the directorate was staffed mostly with people drawn from academia. These personnel choices were in all probability driven by the staffs of the existing directorates whose characteristics closely matched that of both their clientele (the academic community) and their rhetorical function (foster basic research). One result of the staffing decisions was that the working bureaucracy of RANN regarded the academic community as the directorate’s most important market (the source of significant incentives and disincentives) and failed to satisfy either their designated clientele (users of the applied research) or their funding source. An evaluation commissioned by a new director confirmed this. This evaluation eventually resulted in a sharp budget reduction that led to a major reorganization in 1977 to 1978. It is too early to tell if the incentive structure has actually been altered.
Those parts of the governmental organization that are farther from the source of authority perceive the incentive structure differently than those at the top. At the point of direct intervention, the receivers of service (and others who inhabit the immediate environment of the direct intervention) are sometimes the providers of both incentives and disincentives. These are the people who have the most direct contact and who applaud or hiss, make the phone calls, picket, and may instigate action by members of Congress. The stored memory of those parts of the organization closest to the direct intervention may indicate that the organization should propitiate its clientele. [Occasionally, the direct-intervention domain is quite insulated from its clientele and is, at the same time, well removed from its money-generating market. In those cases, the intervenors will usually try to please their bureaucratic superiors (who control promotion) or will sometimes simply please themselves.]

Thus, in the governmental organization, usually at least two tides are running—one carrying the organization in the direction of pleasing the source of funding, the other toward pleasing those involved in the direct intervention.

This schism within governmental organizations has a significant implication for evaluation design since those nearest (particularly those who are in) the direct-intervention domain will be most interested in, and most likely to be able to use, evaluation information that relates immediately to the operation (for example, “Youngsters in the summer camp for undernourished children are receiving Kool-Aid as a lunch beverage instead of the milk intended”). Those who are nearest the source of authority will want, and be able to use, something more global (for example, “3,031 children participated in the summer-camp program for undernourished children” if the program did not meet its stated goals; “2,973 of the 3,031 children participating in the summer-camp program for undernourished children showed a significant decrease in symptoms associated with malnutrition” if it did).

Refer back to figure 1-1 for a moment. The evaluation-information part of the loop splits, with some information going directly to the direct intervention, the rest funneled through those in charge. This implies that both kinds of information have to be gathered. In some cases, it also means that two reports should be written. (Both reports normally contain all of the information. The differences are in emphasis and language.) In other cases, particularly when there is a danger of wiping out an otherwise good program that just needs a little fixing, the information to the direct-intervention domain might better be given informally.

**Effects of Organizational Incentives on the Individual**

In the course of their work, everyone makes many small decisions or takes action that causes results. Through some mechanism or another,
either single or cumulative positive or negative rewards ensue from the results. This is a continuing process.

Consider a machinist in a manufacturing organization who runs out of feed stock for making a particular part. If the exact stock is unavailable, the machinist may select some nearly equivalent feed stock and continue making parts. This small decision may be ignored, rewarded, or rejected by the larger organization or by the customer for the machine containing the part. The machinist stores (almost unknowingly) some information about this transaction. What is stored in the machinist’s mind will be quite different depending upon the outcomes, which could include praise from production management for meeting the quota while out of material, no notice ever taken of the decision, all substitute parts rejected by quality control, or three months later, identification of that decision as the source of a series of failures in customers’ machines due to the substitute parts. If faced with the same situation next year, the machinist’s response will vary depending upon the prior feedback.

Similarly, a social worker facing a problem in counseling selects an approach or technique, carries it out, and obtains some result. Further beliefs about the efficacy of that application of the technique can come from further information from the client, co-workers, the supervisor, or perhaps information about other people who are using the technique. Again, the way the information is internalized will depend upon the information received. Did the client suddenly recover his mental balance, or did he leap from the window? Was there no further feedback at all, or did extensive publicity for the agency and the worker result from the action?

A governmental worker at the upper bureaucratic level, looking around at colleagues’ progress, on the one hand may see few rewards and major dangers in participating in the lengthy, tedious, and difficult process of implementing a program and finding out whether it works. On the other hand, people who provide a quick rhetorical response to requirements from the secretary’s office or from the funding subcommittee may be clearly seen to move quickly ahead. Some years ago, one of the authors was asked to help in the planning of one of the let’s-reorganize-the-executive-office spasms that periodically shake Washington, D.C. I was told that the major effort was to create a structure that could provide quick response to questions from the president and was shown an organizational plan that, if implemented, indubitably would accomplish that purpose. Acknowledging the fact that a quick answer could be produced, I pointed out that no mechanism existed for producing a right answer. The result of that observation was that I was not invited to participate in any more planning meetings. The authors are still not sure whether, for me, the cessation of invitations was an incentive or a disincentive.

Often, accurate, honest responses to requests for measurement of pro-
gram progress seem to place agency funding in jeopardy. Such outcomes are likely to be internalized. What seems to be at work are responses to signals from compelling markets that are far from (or even at odds with) the intended service market of the agency.

Many proposed management systems for use in government seem to hypothesize a situation much like that in figure 8-1. A selected market \((M_0)\) is used to determine acceptability of performance standards \((S_0)\). An individual takes action and the outcome (or set of outcomes, \(O_0\)) is measured. The deviation \((\epsilon_0)\) from the standard is observed by the individual who subsequently adjusts actions taken in order to produce outcomes that more nearly meet the standard.

In reality, the situation is often more like that in figure 8-2. An individual takes action and the outcome is measured. This outcome, however, is compared to the standards of several markets \((M_0, M_1, \ldots)\). (These markets represent groups such as client groups, funding sources, newspaper/television powers, and so on.) These standards are often quite different from one another, and the importance \((I)\) of each is ranked within the framework of each individual’s store of knowledge. Often, the various error signals \((\epsilon_0, \epsilon_1, \epsilon_2)\) do not reach the individual right away (or at all). However, if the market is powerful and the error large, these additional error signals may descend with some force. If so, they are added to the store of knowledge and condition further behavior. In other words:
Organizations and individuals receive information via several feedback loops.

Organizations and individuals rank the importance of the information sources—that is, markets—according to their perceived ability to dispense or withhold incentives and disincentives. (Note that these markets include individual and organizational value systems and that incentives and disincentives may be psychic. Despite our use of cybernetic terminology, or jargon, we do not conjecture an amoral world.)

Adjusted actions are influenced by some balancing of all the error signals received from the several feedback loops and comparisons and their perceived importance.

Thus, the way an organization responds to the results of a given evaluation—whatever the quality of the work—will depend in large measure on the organization’s previous experience with responses to other evaluations and to other performance-feedback pressures and experience.
To summarize, a shared store of knowledge, distributed throughout the organization, is a primary factor in determining if and how an evaluation will be utilized. Organizational incentive structures, by consistently rewarding or punishing certain behavior, build up in individual members a store of experiential knowledge that shapes future behavior. Because of this, some evaluation approaches and some means of using the information produced may be much more effective in one organization (or one part of it) than in another. The designer must consider this in the design if actual use of the evaluation information in support of purposeful behavior is considered part of a successful evaluation effort. Overall, in government operation, people high in the organization will be more oriented toward the funding source, often a subcommittee, than toward the direct interventions performed.

**What to Look For**

The evaluation designer should attempt to determine where parts of the organization that are expected to make use of evaluation information look for incentives and what standards they may have internalized. Those parts that look toward the source of authority tend to share a common context with that source. That context is almost always rhetorical. The assessment measures and comparison standards that these high-level people will insist on will often be what they believe are acceptable to the source of authority and funding (for example, Congress, the school board)—measures, comparisons, and results that they can take to Capitol Hifi, for instance, as ammunition in the continual battle for survival or increased funding. Improvement of program implementation may not be the issue they face first (see chapter 6).

Thus, high-level staff at NIMH may be far more accepting of measures of the decline of enrollment in state mental hospitals than in measures of mental-health improvement by patients of CMHCs. With the crime rate rising, high-level LEAA officials were much more interested in measures of improvement in criminal justice-system capability (half of their charter) than in crime rates (reduction of which is the other half of their charter). During the time that Congress was intent on using public-service employment as a rapid antidote for unemployment and inflation, high Labor Department officials were much more accepting of measures of public-service jobs filled and the happy results for those employed rather than on measures of the effect on local unemployment and inflation. This is not to get into questions of right and wrong but simply to highlight the acceptability of certain measurements because of the likelihood of funding-source use of the information after it is produced.
Conversely, those parts of the organization that look to their clientele or service activities for feedback and incentives tend to share an operational context with that clientele or those activities. The assessment measures and comparisons that they will insist on (or are more likely to use) usually take two forms. In the first instance, they will want information that they believe is acceptable to those immediately above them. However, they will also want measures that can be used as guides to improve actual program performance—if only to remove the pressure put on by their clientele. In other words, they may be prepared to accept evaluation measures that deal directly with the reality of their interventions (although sometimes not, if the information is also to be passed to those higher up). Thus a mental-health center itself might accept measures of what becomes of its patients or of the effect of its activities on the mental-health problems of its “catchment” area. The local law-enforcement person may want to know if the special patrol really did reduce crime.

Both survival and program improvement are healthy concerns of any governmental organization. Neither can be ignored with impunity. Keeping the subcommittees in the Congress happy and under a warm blanket of rhetoric may be sufficient to maintain the organization’s viability. However, the Congress is an organization in its own right and one of its major markets is the citizenry, which is comprised of some of the very same citizens that make up the agency’s clientele. If an enormous number of citizens are offended, Congress begins to hear about it; and then the agency hears about it. Even if a small number of citizens are offended—and if those citizens happen to be part of an important subcommittee member’s district—the agency will hear about it. The feedback paths may be long, tortuous, and filled with random noise, but they are too powerful to be ignored.

Thus it is that even those in the upper reaches of the those-in-charge domain have some interest in program and program improvement. Also, since even the most program-oriented members of the direct-intervention sector realize that the program depends on the good will of Congress, they have some interest in the happiness of that organization. Both those in charge and the direct intervenors will give a little on measures and comparisons. The questions are ones of emphasis and context.
Why Appropriate Measures and Comparisons May Meet either Acceptance or Resistance

Earlier, we discussed learned behavior in terms of being appropriate or inappropriate, pointing out that as long as internalized appropriate behavior does not cause too much trouble, it tends to be repeated. However, some behavior that has worked for people or organizations becomes inappropriate over time to the point of being downright unhealthy. This may be caused by changes in their own situations or in the environment. Individuals often run into the former problem when moving to a new environment. This is not uncommon, for instance, when industrialists move from business to government. During the Eisenhower administration, for example, Defense Secretary Charles Wilson remarked publicly that “What is good for General Motors is good for the country.” That is a perfectly appropriate thing to say to the stockholders of General Motors. Coming from the Secretary of Defense to the entire nation, however, it was a public-relations disaster.

A later defense secretary, Robert McNamara ran into the second problem. He brought with him to defense a penchant for quantification and numbers that served him well in the first few years. After that, however, people with special points to plead simply made sure that they purchased or created a sufficient quantitative scheme to show that their answers were right. Conditions had changed in the secretary’s immediate environment. Subordinates had internalized his approach as a new method of debate and had begun to engage in creative quantification.

In the case of organizations, previously healthy (or appropriate) behavior patterns can become unhealthy as the political climate changes or as congressional attention is focused on particular program areas. In these cases, if the organization keeps repeating its no-longer-appropriate responses, the incentive structure will slap it hard and often. Even then the old organizational response—internalized by large numbers of dispersed people—may continue to the point that the organization appears to have a death wish. On occasion the wish is fulfilled.

The evaluator who has explored an organization from one end to the other can often sense situational changes (or impending changes) long before members of the organization at particular locations. The discussions of assessment measures and comparisons raised by evaluation designers can do much to reinforce healthy behavior patterns and discourage unhealthy ones or vice versa—avoiding trouble or virtually looking for it.

An example of unhealthy reinforcement can be drawn from the Vietnam war. One of the process measures chosen for evaluation of helicopter units was the number of sorties flown. It was to be one of the bases on which units were rated and promotions given. This was a convenient measure that could be pointed out in Washington whenever
Congress wanted to know what our helicopters were doing over there. It was a serviceable measure that could be used to justify logistical operations, supply purchases, and a whole host of matters that Congress found interesting. Unfortunately, it was also a measure that reinforced unhealthy behavior patterns at the direct-intervention level. In order to score well—that is, raise the sortie rate—commanders could do three things: (1) they could send their helicopters on unacceptably dangerous sorties, risking life and helicopters; (2) they could send their helicopters on short sorties far from the war, thereby wasting fuel and putting pilots and helicopters temporarily out of reach when needed but piling up sorties with little risk; or (3) they could ignore the measure, provide the support for which their units were intended, and pray that the fortunes of war would reward the just.

Much the same thing happens in social programs. In manpower training programs, for example, number of placements is a common assessment measure. Direct intervenors, in order to score well, have been known to keep lists of high-turnover jobs in reserve—car washes are typical. If skillfully done, a direct intervenor might get as many as six placements a year from the same job in a particularly odious car wash. That is unhealthy reinforcement. Meeting the measure does not advance the goals of the program.

Comparison standards work very much like the thermostat in our furnace example. The selection of a standard (or standards) ultimately affects how the process runs. If evaluators expect their work to be used and to result in progress, then the selection of assessment measures and comparison standards must in-elude an informed attempt to accommodate the needs and desires of different parts of the organization. An evaluation that employs appropriate, well-thought-out measures will tend either to reinforce healthy behavior or to alter presently unhealthy behavior. Building an evaluation system or procedure that will reinforce and be reinforced by stored attitudes and behaviors is much simpler than building one that will both meet resistance and require changes throughout the organization.

The choices of organizations in responding to evaluation information are conditioned by the distributed beliefs of people about the effects of the action upon themselves, their peers, others in the organization, and others outside. In a sense, a pattern of consistent rewards and sanctions might be thought of as training the organization over a period of time by increasing the proportion of stored experience that leads to a particular set of outcomes for the organization. If a governmental organization has carried out a successful pattern of obtaining its funding without ever knowing accurately or in detail what it is doing and how performance is coming out, then serious evaluation will not come easily to that organization. Many higher-level members of the those-in-charge domain are so conditioned. If it appears that the evaluation system being designed
must overcome a resistance rather than be reinforced by existing beliefs, then much more organizational design and leadership prestige must go into the work for it ever to become successful.

Notes


This Page Intentionally Left Blank
Over the years it has become apparent that many programs of government did not achieve their objectives either because expectations were patently unrealistic, because the program was not implemented as designed or expected, or because the underlying reality of the world was different than the policymakers assumed. These recurrent bases for program failure have plagued the federal government in diverse areas such as budgeting, social programs, economics, defense, energy, and regulation.

In chapter 4, we reported the results of one program evaluation. There, projects had not been implemented as planned or expected, and they apparently never affected the system they were intended to. A reasonable question, for which we could find no reasonable answer, is: Why did anyone attempt to measure the broad effects of a program that did not exist?” In chapter 9, we cite several larger-scale examples for which legions of analysts did not see it as their responsibility to examine reality carefully. The resulting disasters may be laid at least partially at the door of these analysts. Between us, we have examined nearly 1,000 reports, evaluations, cost/benefit studies, and program, policy, or research analyses for high-level officials. These reviews have led us to the sobering conclusion that reality very often is not accounted for when activities are planned, putatively managed, and ultimately evaluated. The approach outlined in this part is an approach to examining reality and to distinguishing between information based on fact and information based on opinion or even pipe dreams.

To return to our furnace company’s black-box model in chapter 2, the performance of an intensive evaluation without checking to see if the program is in place is tantamount to sending a service crew to the house without first determining whether the oil ever got to the tank (or if there was even a furnace in the basement).

It would seem obvious that if the direct interventions of a program have not been implemented as designed, then it is not possible to assess the effectiveness of the program as designed. Nonetheless, evaluators continue to attempt such evaluations, as demonstrated by the example in chapter 4. Those programs are not evaluable. The process of evaluability assessment (EA) was initially developed to weed out candidates for evaluation that could not reasonably be expected to achieve their objectives—that is, EA is a process carried out between the time when an activity becomes a candidate for evaluation (including congressional oversight) and the time when the evaluation is finally designed. The original and primary purpose of conducting an EA is to increase the probability that the eventual design and performance of an evaluation will
produce usable, used results. A secondary, and possibly even more useful, purpose has emerged over time—namely, the process itself has proved to be an excellent management tool in that the information produced through the EA process is often enough to tell those in charge what they need to know about their program in order to take effective remedial action.

In essence, the process of EA is a systematic way of answering the most basic questions first: What was to be done? What activities are in place and functioning? Is there reason to expect that the program outcomes or impacts postulated have, in fact, occurred? Does the world of the direct intervention resemble the world expected? What can be determined, in what sequence, and at what cost?

The chapters in this part describe the process of EA. Chapter 9 provides an overview of methodology. Chapters 10, 11, and 12 detail the three classes of models (testable, equivalency, and evaluable) used during an EA. Chapter 13 explains how a series of preplanned reviews permit those in charge to purchase only as much information as they believe will be useful. Finally, chapter 14 discusses the wide range of possibilities for gathering EA information.
The world contains many different commodities—some plentiful, some scarce. With some notable exceptions (for example, skunks in a high-rise), the scarce commodities are highly desired and costly. Usually, questions are in plentiful supply, but accurate and valid answers are scarce.

Hiring evaluators represents the purchase of the services of a particular team of answer brokers. The evaluation team is charged with spending a portion of the organization’s resources to systematically acquire information that can be used to guide further actions of the organization.

Any team of honest brokers will admit from the start that the universe of possible questions can be separated into three categories:

1. Questions that can be answered at reasonable cost,
2. Questions that can be answered at unreasonable cost,
3. Questions that cannot be answered at any cost.

The information that emerges during an EA serves to sort the questions into these categories so that the potential users of the information will waste neither time nor money in pursuit of answers that are infeasible or impossible to obtain.

An EA is built on four cornerstones:

1. The construction of two families of models—namely, testable models based on information derived from descriptions and equivalency models based on information derived from observation;
2. Comparisons and reconciliations within and between the two families to produce an evaluable model, often the basis for immediate action and always the basis of evaluation design;
3. The construction and use of functional models (see chapter 3) to display the relevant structure and flows of both the described and observed activities of interest;
4. A phased approach to the entire investigation that permits those in charge to make sequential purchases of information.

The remainder of this chapter contains an overview of EA, and the subsequent chapters examine the process in more detail.
Testable and Equivalency Information

EA employs two distinct classes of information. First, it uses testable information; second, it uses equivalency information. The EA process is shown in figure 9-1. Both classes of information are gathered—usually simultaneously—from many sources, as shown on the left side of the figure. The testable information is accepted uncritically and combined to produce the two models within the testable family—namely, the testable logic model and the testable functional model.

The logic model is often framed as a series of if-then statements that outline the framework of expectations—for example, If we spend $1 to set up child-health projects, then we will return unhealthy children to a state of basic good health, and then all enrollees and parents will maintain good health practices throughout life. The logic model, which displays the programs’s concepts and expectations, is one form of the goals-and-objectives statement so familiar to those in charge of government programs.

The testable functional model, always a flow and function diagram, describes how various stakeholders believe (or say they believe) that the activity of interest (usually the direct intervention) has been implemented. Figure 9-2 is a rudimentary testable functional model of a child-health program. (In actual EAs the model would show much greater detail.)

Notice that even within a family of models, there may be important discrepancies between the individual models. In the case illustrated, for instance, the logic model concludes with a triumphant “and then all enrollees and parents will maintain good health practices throughout life.” The testable functional model, however, shows only health education as an implementing function for lifelong health maintenance—no long-term follow-up of patients who have completed the program was contemplated.

The models derived from testable information display the belief structure on which those in charge base their questions about the activity of interest. In the child-health program, for instance, those in charge might well be interested in aspects of the enrollment process (How was the door-to-door canvass accomplished? How many people were canvassed?). They might be interested in the outcome of the health-education element (What do the graduates know about dental care?) or the treatment element (How well do the assigned doctors do compared to the patients’ own doctors?). They also might be interested in longer-term impact (Are the individual children still healthy after five years? Is the community, as a whole, healthier after five years?). It is important to recognize that these questions are based on a belief structure, on the testable class of information. Whether or not the questions are answerable depends not on belief but on reality.

The equivalency functional model is based on direct observation of
Figure 9-1. Basic Distinctions Used in Evaluability Assessment
Figure 9-2. Testable Functional Model of Child-Health Projects
the activity of interest, and unlike the testable models, the evaluators are responsible for its accuracy—that is, the evaluators must ensure that the model accurately depicts the salient structure and flows of the operating system. The equivalency functional model is the most important member of the family of equivalency models. This family also includes a summary statement that has been derived from the equivalency functional model. Figures 9-3 and 9-4 are equivalency functional models that depict how two different child-health projects were actually implemented.

The functional models, derived from equivalency information, display the operational reality of the projects. Based on that reality, a summary statement, in a form similar to the testable logic model, can be extrapolated. In the case of child-health project I (figure 9-3), the summary statement would read: If we spend $X to set up a child-health project, then we will return unhealthy children to a state of good health. In the case of child-health project II (figure 9-4), the summary statement would read: If we spend $X to set up a child-health project, then local health professionals will treat unhealthy children for a year.

Note that the testable logic model and the equivalency summary statement are derived in totally different ways. On the one hand, the testable logic model is constructed before the testable functional model—as a concept statement, it is the first stage in planning a program. The equivalency summary statement, on the other hand, is constructed after the equivalency functional model is roughed out. It is the evaluators’ extrapolation of what the concept must have been based on activities that have been implemented. It is a summary statement of reality.

**Comparisons and Reconciliations within and between the Families of Models**

In the example given, the comparisons between the two families of models are striking. Not only are the structures and functions of the actual projects different from the believed functions and flows, but also the logics are markedly different. Further, as noted earlier, there is a serious discrepancy between the logic and functional models within the testable family.
Figure 9-3. Equivalency Functional Model of Child-Health Project I
Our example is an adaptation (for simplicity’s sake) of an actual national child-health program—one with twenty-nine projects nationwide, the projects each a little different from one another and from the expectations of those in charge. At this point, it is helpful to walk through some of the comparisons to see how questions are sorted into the three categories (questions that can be answered at reasonable cost, at unreasonable cost, not at all) and what actions might be taken prior to the construction of the evaluation design.

In order for a program to approach the expectations for it, each outcome contained in the logic model must have an implementing function or a set of implementing functions. (Early in a program’s history, before the direct intervenors are hired and dispatched, the construction of these models, or some surrogate for them, comprises the program-planning phase of the operation.) As we saw earlier, one outcome of the testable logic model, that all enrollees and parents will maintain good health practices throughout life, had no tangible implementing function—unless you are willing to believe that attendance in a brief health-care course leads to the maintenance of lifelong good health practices. No long-term follow-up reinforcement was envisioned, nor was any long-term tracking planned. Thus, the question, Are the kids still healthy after five years?, is impossible to answer if the projects have been implemented in accordance with the beliefs of those in charge. Further, any evaluation design that attempted to answer that question would have incorporated a type III error—measuring something that was not there (see chapter 4).

The early recognition of the discrepancy could be the basis of several courses of immediate action by those in charge. One course would be to check to see whether the field projects had noticed the discrepancy and had inserted—on their own—an appropriate implementing function. If that had happened (such field initiative is not unknown), those in charge would simply have changed their belief structure to incorporate the long-term function into the testable functional model. A second course of action would be to issue a directive telling the field units to implement a long-term function and then to change the belief structure.
underlying the testable functional model. A third course would be to change the expected outcomes of the program and to drop the long-term element from the logic model. In any event, the models within the testable family would be reconciled.

There also must be comparisons and reconciliations between the testable models and the equivalency models. First, the testable logic model and the equivalency summary statement are compared. This comparison involves the initial concept statement embodied in the testable logic model and the summary equivalency concept statement(s). As in the within-family comparison described in the last section, this comparison could be the basis of immediate action—including calling the whole evaluation off.

Assuming that the EA continues, the two functional models are compared. In our examples, neither of the actual child-health projects incorporated the expected health-education function—it was found to be too expensive to implement on-site—nor did either project provide long-term treatment and/or tracking.

In child-health project I (figure 9-3), the enrolled child was restored to basic health—however short or long a time that took—and then returned to the target population. Long-term follow-up and care were not provided. Child-health project II (figure 9-4) used entirely different methods of recruitment and examination. In addition, the enrolled child was not necessarily restored to basic health. However, one year’s worth of continuing care was given to all enrollees diagnosed as not OK during the initial examination.

It is evident that any evaluation design that incorporated all of the measurements suggested by the equivalency functional models could easily contain a number of type IV errors—that is, measuring something nobody is interested in (see chapter 4). In child-health project II, for example, the time spent per enrollee for the initial screening and examination might not interest anyone, particularly if those in charge had decided that nurses and aides, not the project director, should do the screening.

Like the comparisons mentioned earlier, the recognition of discrepancies between the testable and equivalency functional models can lead to immediate action by those in charge. The options for action range from a simple change in the belief structure underlying the testable models to a major overhaul and close monitoring of field projects—along with concomitant changes in the equivalency models. The reconciled functional models form the evaluable model—the model that is an accurate representation of that part of the activity of interest to those in charge. The evaluable model is used as the basis for discussion of the expectations, questions, costs, and evidence that is acceptable to those in charge.
Sequential Purchase of Information

There is seldom enough equivalency information to completely plan an evaluation when one is first considered. As we observed, each time the models are compared, several options are available to those in charge. If no serious discrepancies exist among the models, those in charge may simply continue to purchase the EA information as originally scheduled. When the models differ, however, the range of possibilities is virtually limitless. In our example, for instance, those in charge might have decided that more information was needed (Are any of the twenty-nine projects implementing as expected? Scan the field and if one does conform to design, concentrate the modeling efforts there); to change their belief structures (The testable functional model better include the diversity of implementation, and the evaluation should forget about measuring the effects of health education); or to change the project implementation (We will send a directive stating that nurses have to do the screening. Postpone the EA and monitor the projects for a few months so that we are sure they are conforming).

The parenthetical examples illustrate the three categories into which questions are sorted. The first quote is probably one that can be answered at reasonable cost, and if such a project existed, an evaluable model could be constructed with relatively little pain.

In the second parenthetical quote, the statement, “The testable functional model better include the diversity of information,” implies that those in charge will have some questions about how the efficacy of the implementation methods compare. Those questions can probably be convincingly answered, but only if great care is taken in selecting comparison groups and controlling stray variables. It is entirely possible that the cost of answering the questions would be unreasonable. The rest of the second quote “forget about measuring the effects of health education,” demonstrates the recognition that some questions cannot be answered at any cost.

The process of sequential purchase based on developing equivalency information is detailed more fully in chapter 13.

Two Ways to Foul-Up an Attempt at Oversight

The EA process of comparing and reconciling information based on descriptions against information based on observation of reality—by whatever name the process is called—has been used increasingly by government evaluators over the last ten years. The reason for the increasing use of the process is evident: It allows the evaluators and the people who hired them to come to grips with the problems engendered by the gaps between rhetoric and reality. Those problems are commonly found in government operations and, when ignored, lead onward to
embarrassment, failure, or national disaster.

On the executive side of government, the U.S. Department of Commerce recommends the process as an “appropriate evaluation procedure.” On the legislative side, those involved in government oversight have shown considerable interest in and knowledge of the process. The GAO, in particular, has propounded the process.

In his excellent report to the congress on congressional oversight, the comptroller general of the United States reported:

Requirements for Matching Program Information to Congressional Oversight Needs

It is important that evaluation measures and comparisons reflect both the legislative or policy intent and the actual program activities being carried out. Most of the oversight questions to be answered by program-results reviews (i.e., those that concern progress toward goals) come from the Congress and policy-level personnel in the agency. On the other hand, at the program delivery point some “real process” is being carried out on a day-to-day basis. In a program evaluation study, it is this real process and its effects that are measured to produce answers to questions concerning program outcomes and impacts. Evaluators are usually the people who must make these real measurements and convert the measurements into answers to oversight questions. The evaluators are among the first people (and occasionally the only people) who encounter the problem of extracting, through an actual measurement of concrete, real-world situations, the answers to questions which have been shaped by abstract statements in the political or policy world. The evaluators must determine from the rhetoric of policy exactly what was intended and then make actual measurements to see if it occurred.

Experience has shown that it is tremendously difficult to specify accurately in advance, even in ongoing programs, the correct obtainable measurements before the implementation, program monitoring, evaluation, or oversight effort is begun. The solution to this problem seems to lie in proceeding iteratively and encouraging sequential discussion and agreement.

At least two serious risks occur when the basis of oversight discussions is not specified iteratively. Consider first the risk of not understanding the actual process occurring at program delivery points when an evaluation is structured. If the actual process is not well understood and the evaluation is designed directly from the abstract policy descriptions of what should happen, then the design of information collection simply may not “fit” the actual operation of the department or project. After much effort, time, and measurement, the evaluators may only be able to show that the world is quite different from what it was thought to be. In the public arena this will often be hard to distinguish from program failure. A misdesigned evaluation or oversight effort, even though it produces accurate answers from an effective program, may adversely affect a program simply because the rhetoric about the program had been unrealistic. This is a serious danger resulting from the
use of faulty, simplistic evaluation designs determined exploration of the actual implemented program.

There is a second risk that works in exactly the opposite way—the evaluator may design the evaluation with an eye only on the direct activities. The evaluator may come to understand the actual process very well. The evaluator may, again at some expense, time, and effort, make a series of careful measurements from the actual operation. But in this case, when the attempt is made to translate the measurements into information for the legislative and executive debate, the evaluator may find that none of the things that figure in that debate has been measured. A perfectly valid evaluation of the actions taken may be performed. But if the information is unrelated to the issues of the policy debate, it is irrelevant. The oversight debate may take place entirely in terms of something that has not been measured.

Much has been learned by evaluators over the past ten years about methods to avoid the two kinds of damaging results discussed above when attempting to conduct oversight, how to avoid wasting the evaluation or oversight resources, and how to make the evaluation product more useful to the decision makers. In most cases these insights involve the design and conduct of a process rather than the issuance of specific guidelines, directions, or standards. Despite this accumulating base of knowledge, some planning and management systems introduced in the last few years within agencies have created a plethora of unread (and often unreadable) material and very little discussion and agreement on management and evaluation measures.

The purpose of this report is to suggest ways in which the Congress can avoid this mistake. The key to obtaining maximum usage would seem to be to develop a process that produces discussion, agreement, and oversight, with a minimum of paperwork.

The report goes on to lay out in detail the problems that the distinctions of EA bring to the surface and some of the cures for them. In this section of the overview, we briefly illustrate the kinds of foul-ups mentioned in the comptroller general’s report. A more-detailed discussion is given in chapters 10 and 14.

The first foul-up is that of designing an evaluation using only testable information. As we discussed before, this can lead to type III errors (measuring something that does not exist) and money wasted. These errors are often occasioned because those in charge hold unrealistic expectations for the program. There are a number of variations on this theme and a number of reasons for their occurrence. In some cases, expectations and plans may have been formulated before the budget was allocated. With the advent of a disappointing allocation, those in charge might scale down their activity plans without similarly scaling down their expectations. For example, a logic statement—If we spend $X to set up child-health projects, then we will return unhealthy children to a state of
basic good health, and then all enrollees and parents will maintain good
health practices throughout life—may be unrealistic if the program
allocation falls short of expectations. Or, bureaucratic lead times and
political factors may contribute to a serious discrepancy—for example, an
unexpected hold-up in processing even a small contract could result in a
military command’s not being appropriately equipped to perform an
expected operation.

Another typical cause of unrealistic expectations occurs in planned
interagency projects where all or most of the agencies except the
designated lead agency pull out before the project gets off the ground.
The temptation to go it alone is almost overwhelming. A case in point
was a planned integrated social-service-delivery program in which several
agencies were expected to deliver their wares (for example, employment
information or senior citizen recreation) through a variety of
related communications technologies. The goal of the program was to
demonstrate the economic feasibility of aggregating uses of the
technologies. Even before serious planning began, all of the agencies,
with the exception of the technology-oriented one, decided against
participation. At this writing, the logic of the program is unchanged.
However, without contributions of money and expertise by the other
agencies, it is unrealistic to expect that either economic feasibility or
aggregated service delivery can be demonstrated. A more-realistic
expectation would be that the technology will operate smoothly.

The logic of intermediate events may also render expectations
unrealistic. For instance, the money allocated may appear sufficient, but a
large portion of it may be drained off by overhead-type activities before it
reaches the intervention. Most typically this occurs when allocations for
planning are not made separately from implementation. It is possible to
eat up so much money during the planning phase that the effect is the
same as if the allocation had been obviously insufficient.

The second foul-up—working only from observed activities—can
lead not only to type IV errors (measuring something that no one is
interested in) but also, as the comptroller general pointed out, to
irreparable damage to a worthwhile program by failing to measure things
of great interest. In our example, for instance, a perfectly valid evaluation
could measure the differences in efficacy of the various recruitment
mechanisms (a measurement of only mild interest) and at the same time
miss an important element in the policy debate—perhaps, for example,
whether children in given socioeconomic groups were being restored to
good health. For lack of that kind of information, a program could be lost.

Consider some classic governmental cases of testable statements
versus equivalency. In the 1960s, Robert McNamara made his famous
“light at the end of the tunnel” statement about the Vietnam war backed
by endless quantitative analyses of evaluative data. Neither the
equivalency functional model of the war nor the sources of the testable
information was attached to these evaluations. The actual war dragged
on; and so did complex evaluation systems that were not rooted in equivalency. A relatively simple EA could have avoided this mistake.

Also in the 1960s, Mayor John Lindsay announced that, through new and advanced financial techniques, New York would easily be able to finance widespread additional services. Legions of analysts were assisting him. However, the equivalency model (on which the analysts did not comment) showed that long-term borrowing was being used to finance short-term operations. Not long after the major’s announcement, it became obvious that New York was tottering on the edge of bankruptcy.

The Department of Energy (DOE) was founded in 1977 to produce (among other testable goals) energy independence. At this writing, their so-called sunset evaluation (legally mandated in the law as Title X of Public Law 95-91, The DOE Organization Act of 1977) is beginning to get underway. It appears that the DOE so far has failed to approach the testable expectation. Equivalency information, used earlier instead of endless economic and computer projections, might have brought these issues to closure and, in fact, might do so yet. Since the DOE had a policy of not performing extensive program evaluation, few people saw fit to validate an equivalency model of the situation and examine expectations against reality.

Over the last four years, one of the authors (Nay) expanded his work into regulation. The testable intent of many regulatory policies must be carefully and painfully sorted against the equivalency of the regulation and the effects on the actual regulated industries before sensible evaluation of a regulatory thrust can be attempted. Endless delays in carrying out quite reasonable regulatory mandates can lead to extensive economic damage to a regulated industry.

A combination of the political process, size, complexity, and popular beliefs (even among specialists) conspires in government operations to make the testable information difficult to use. The distinction between testable information (and expectations) and equivalency information has thus become too important to ignore.

Presidents, as well as the Congress and officials, are frequently involved in recharacterizing testable expectations. Addressing a joint session of Congress on 29 April 1981, President Reagan said:

"Probably the most common misconception is that we are proposing to reduce government revenues to less than what government has been receiving. This is not true. Actually, the discussion has to do with how much of a tax increase should be imposed on the taxpayer in 1982."

"Now I know that over the recess in some informal polling, some of your constituents have been asked which they would rather have: a balanced budget of a
tax cut. And with a common sense that characterizes the people of this country, the answer, of course, has been a balanced budget.

But, may I suggest, with no inference that there was wrong intent on the part of those who asked the question, [that] the question was inappropriate for the situation. Our choice is not between a balanced budget and a tax cut. Properly asked, the question is: Do you want a great big raise in your taxes this year or, at the worst, a very little increase with the prospect of tax reduction and a balanced budget down the road aways?

Only time and careful investigation will reveal the equivalency against which the corrected statement of expectations of the new president about the 1982 budget must be compared and tested.

**Big Organizations Make It All Harder: Hierarchies and Discontinuities**

In chapter 5, we began to lay out the multiple levels of organizations usually encountered in governmental operations. Chapter 14 displays in some detail the many sources of information that this creates. The long line stretching from legislative action through agency operation and into society at large provides a dimension that must be accounted for in successfully pursuing either oversight or evaluation work. The complicated organizational hierarchies of programs, regulation, or defense can affect analyses or operations in several ways.

Many foul-ups can occur in the linkages between the legislative source of authority and the direct intervention in any government activity. The more layers in the those-in-charge domain, the more chances for a glitch, or malfunction, to get into the system. A glitch can creep into the line of authority, where regulations and directives, issued at different levels, can gradually bend the initial intent out of shape. The bend can be inadvertent as implementation plans imperceptibly drift from intent as they are passed from level to level, rather like the whispering game of children (language differences among levels pose problems for more people than evaluators). Or, the bend can be deliberate— the policy level may simply not like the intent of the legislation, or the working bureaucracy may be asserting itself, which typically happens when the direct intervenors heavily weight the desires of their clientele (their major market) at the expense of the desires of those in charge. This was the case with a tuberculosis-testing project carried out in a midwestern city some years ago. Located in the skid-row area of the city, the direct intervention was expected to identify carriers of tuberculosis. Instead, the direct intervenors responded to the perceived needs of their neighborhood and turned the direct intervention into a clinic for alcoholics and drug addicts. Eventually, an intensive evaluation disclosed that the number of tuberculosis cases identified fell far short of the number expected. This
result provided the first clue that something was amiss in the direct intervention. (When the clinic director was asked why he had ignored his charge, he explained that alcoholics and addicts are particularly vulnerable to tuberculosis, that he really had not ignored his charge—he had just expanded it a little.)

Yet another type of glitch in the line of authority is seen in cases of the missing link. Here, directions and/or funds authorizations may be passed down to empty offices where they await personnel to fill empty slots. This situation occurs most frequently toward the end of an administration, when political appointees peel off to private-sector jobs and, at the beginning of an administration, when the empty offices await congressional confirmations of political appointees.

Glitches can also appear in the chain of ancillary activities, where required enabling activities may not be performed for all of the reasons mentioned before plus the frequent necessity of relying on another agency (for example, the General Services Administration) to implement plans. Chapter 14 covers, in detail, the problems presented by big organizations, their hierarchies and discontinuities.

Synopsis

Different practitioners have adopted slightly differing techniques for conducting an EA as the process has evolved. All of the techniques, however, are built on a framework that consists of:

1. The construction of a family of testable models based on rhetorical descriptions of the activity under consideration;
2. The construction of a family of equivalency models based on the evaluators’ direct observations of the direct interventions of the activity under consideration;
3. Comparisons within and between the two families to determine whether the questions of interest to those in charge can, in fact, be answered during an intensive evaluation of the effects produced by the direct intervention;
4. A phased approach to the purchase of information.

Only if the determination is positive is a program considered evaluable and an evaluation designed. The invention and early development of EA is treated briefly in the appendix.

In the discussions of the models that follow in chapters 10, 11, and 12, it is important to remember that the referent of the models is the direct intervention. That is the crux of what is being described and what is being observed. Of primary interest are the comparisons between the testable and equivalency models of the direct-intervention domain. If those models are closely congruent, the team can move on to an evaluable
model and an evaluation design. If things are not going according to plan, but if the problem can be identified as originating in the direct intervention itself, then the evaluators return with the bad news to those in charge, who may exercise one of their options—namely, stop the evaluation and the program, stop the evaluation but not the program, commission an evaluation based on the testable model, change what is happening in the direct intervention, or change their beliefs about the program. These types of sequential purchases are discussed in chapter 13. In some instances, though, the glitches do not originate in the direct intervention but are, rather, glitches in the links between the source of authority and the direct intervention. In those cases, the comparison between the testable and equivalency models of the direct intervention will provide invaluable clues to the identification of the offending glitch: Did some expected contribution to the process fail to arrive? Did reasonably interpreted guidelines and regulations (issued by those in charge) violate the program’s concepts (formulated by legislation)? Has the nature of the bureaucratic world frustrated a well-intended (or perhaps even needed) intervention? Did the people who formulated the program base their concept on mistaken assumptions about the structure of the environment in which the intervention is taking place?

As such situations become apparent, it is possible to isolate the level of the those-in-charge domain at which the glitch occurred and to construct testable and equivalency models with that level as the referent. A detailed examination of this type (described in chapter 14) will almost always trace the problem (often the result of a drift away from reality) and provide a blueprint for its resolution.

Notes

The family of testable models represents the activity being studied as it is believed to exist by those in charge of the organization. The two testable models—logic and functional—are both syntheses of the various beliefs and show the chains of logic, flows, and functions believed to exist within the direct-intervention process through to the expected impact. Taken together, this family lays out expectations for the program, the believed activities of the program, and suggests the loci and kinds of measurements and the realization of expectations.

The construction of testable models can be wonderfully easy, impossibly difficult, or something in between, depending on the size, complexity, and cohesiveness of the organization and on the bureaucratic location of the buyer of information, relative to the direct intervention. As a rule of thumb, the collection of rhetorical models for synthesis into testable models covers the bottom level of the those-in-charge domain up through one level above the purchaser of the information. Thus, if the buyer of information is close to the direct intervention, the synthesis will not include the rhetorical models of very many layers. If you look back at figure 5-2, the point can be clearly illustrated. Suppose that the principal in the school system commissioned an EA of audio-visual instruction in a given school. The testable models would only reconcile the descriptions of the process and its intended effects given by the principal and the deputy superintendent (and perhaps some key staff members). The deputy superintendent is the principal’s immediate bureaucratic superior. The deputy’s model is of interest in large measure because it will exert great influence over the measurements and comparisons that the principal will be willing to accept if and when the intensive evaluation is performed (see chapter 6).

Contrast the magnitude of that task with the magnitude of the one that would be required had the purchaser of information been the administrator of the ADMHA, who was trying to determine if CAMHCs were effective. In the former case, not only is the number of layers quite small, but also mutually agreed-upon measures of effectiveness would probably not be difficult to attain because standardized measuring instruments and comparisons are in common use. In the latter case, not only are there many levels to consider (the deputy assistant secretary of health through local CAMHC directors plus a few lateral models of interest) but also all of the possible measures and comparisons implied in figure 6-1. An EA conducted for the principal would probably be a breeze; the one for the administrator would in all likelihood be a bear. Keep in mind, as we discuss the two members of the testable family, that on the one hand the construction of the models can be almost as easy as it
sounds; on the other hand, do not count on it.

**The Testable Logic Model**

The logic models (usually displayed as if-then statements) display the expectation that those in charge have for the program. The logic models can be extremely straightforward statements (for example, if I step on a crack, I will break my mother’s back) or quite complicated in cases where bevy of inputs are expected to result in varieties of outcomes and impacts. Usually, the entire chain of program expectations itself represents a portion of the expectations that those in charge have for society. For instance, the democratic weltanschauung holds that a free society rests on an informed and enlightened electorate. This concept leads to the encompassing logic model shown in figure 10-1.

To some extent, the logic models of most school programs display intermediate expectations subsumed by that encompassing model. We can examine some knowledge-transfer logic models within that framework. Suppose that the assistant superintendent for instruction had been approached by a manufacturer of computer-assisted instruction (CAI) hardware who convinced her that the school system should try the CAI devices on a demonstration basis. Together they write a proposal to the state office of education and receive a revenue-sharing grant to initiate the demonstration and evaluate it.

The logic model of the superintendent of schools would look something like the logic model in figure 10-2. There are three things to notice about the logic model of the superintendent. First, it is quite straightforward. Second, it scopes the project—that is, if an evaluation team was to be called in, the team would neither concern itself with aspects of the school system that have no connection to the CAI program (for example, the fourth-grade art program) nor with the place of CAI in the encompassing model (for example, the impact CAI has on the development of a free society). Measurements of things beyond the scope of the logic models are among the type iv errors (see chapter 4).

![Figure 10-1. One Overall Logic Model for Universal, Free Education](image-url)
Third, the logic model does not include well-defined functions. It simply states some basic concepts about the program. Logic models are for the purpose of scoping and communicating. They are not appropriate tools for examining details of structure or cause and effect, which is the role of the functional model. There are many possible functional models (and levels of structural detail within the functional models) for each application. There are many, many logic models for each functional model. To reiterate, logic models are for scoping and communicating, not for analyzing.

Within an organizational hierarchy, each level will see the project from the perspective of its own functions and each will add logic statements of its own. As the various models are reconciled and the inputs and outputs are fitted together, the combined logic model provides a chain detailing the believed inputs and outputs of each group. With the accumulation and reconciliation of more-and-more testable logic, the logic model can grow quite detailed.

The assistant superintendent for instruction, for example, would have a different, complementary logic statement, as shown in figure 10-3. Some people on the assistant superintendent’s staff would probably have other logic models. For instance, the curriculum coordinator’s model might be as shown in figure 10-4; the director of evaluation’s might be as shown in figure 10-5.

At the level of the principal of the school in which the demonstration project is taking place, another consideration might enter the logic—namely, future instructional-staff redeployment. The principal’s logic model is shown in figure 10-6. Notice that B of the principal’s
Notice that B of the principal’s model is an assumed outcome of an affirmative finding for the superintendent’s cost-effectiveness questions (figure 10-2, box B) and that the principal’s C is an expectation of future impact not mentioned by the superintendent. That is a potentially sticky point, however, since the superintendent may have other plans for money saved. The evaluation team must decide whether to bring up the matter during discussions of the logic model, to postpone it, or to declare it beyond the scope of evaluation.

The logic models of the actors we mentioned could each have more boxes. In addition, more actors could be asked to contribute. However, the models shown are sufficient to illustrate the process of constructing a logic model. The combined logic model is shown in figure 10-7.

Notice that even at the logic-model-constructing stage, certain problems such as potential discrepancies in expectation begin to show up, and that certain questions arise about the realism of assumptions—for example, Is there enough money to do all the revising and measuring?

The boundaries of the model imply that an evaluation was commissioned by the assistant superintendent of instruction. If, however, the evaluation was initiated at the state level, then the chain of logic would extend upward, possibly to an assistant commissioner of education who might have a logic model that begins with “If we fund a CAI demonstration project” and ends with “then we will have widespread adoption of CAI in public schools throughout the state.”

One point of this particular example is that if the activity under study is discretely bounded and the relevant actors have a cohesive idea about the goals of the activity, then constructing a combined logic model may be a tedious process but not a difficult one. (For the sake of illustration, we did not include potentially controversial logic models. We might, for
instance, have included the models of guidance counselors with strong opinions about the potentially adverse effects of CAI on the process of preadolescent socialization.)

If the program under consideration is not a discrete, focused project like the CAI example but, say, all of the activities that comprise a school system, both the size of the combined logic models and the difficulties of the task become orders of magnitude larger. A model of that nature beings to approach the philosophic complexity inherent in the encompassing model that ended with the creation of a free society (figure 10-1).

Another clue to the potential difficulties in constructing logic models is evident in the possible impacts of day-care centers that were listed in chapter 1. They were as follows:

- Inadequate child care will be replaced with adequate child care.
- Children will live up to their potential.
- Nutritional and health-care standards of poor children will be raised.
- The net income of the poor will go up.
- Welfare rolls will be reduced.
- Women will be liberated to enter the work force if they desire.

Try to imagine how six different logic models, each ending with one of the expected impacts listed, might ‘be plausibly combined, given the level of funding available for day care at the time when all of these models were rhetorically current (1969). We asked several people to try to construct a plausible testable model. Some very clever people could combine five of the models; no one could plausibly combine all six.
Figure 10-7. Example of a Combined-Logic Model for CAI Program
Some cases will occur in which just the attempt to reconcile the logic models of those in charge will provide a strong indication that the program under consideration is not evaluable.

The Testable Functional Model

For the evaluators, the procedure for getting testable information for a testable functional model is exactly the same as that for getting information for the testable logic model. The evaluators go to each level of the those-in-charge domain and ask the relevant people for descriptions. In the case of the testable functional models, the descriptions will be of the process taking place in the direct-intervention domain. The combined testable functional model, however, will be a synthesis of the various descriptions—this time tracing the believed flow of a unifying concept or concepts (see chapter 3) from the time the unifying concept arrives in the direct-intervention domain through the time it passes beyond the boundary set by the logic model. The functional model displays the believed flow of the unifying concept(s) and the functions that are performed along its path.

In the discussion of testable functional models, instead of reiterating the step-by-step process of combining the testable models, we go directly to the process of selecting the flows of interest. Figure 10-8 is the most general functional model (a black box) of a prototypical direct intervention (adapted from figure 1-4).

Having decided that the unifying concepts of greatest interest are probably knowledge transfer and students, we can begin to fill in the generic designations with specifics that are pertinent to the unifying concept. We start with inputs, contributions, outputs, impacts, and control/comparison groups that have been described and seem to fall within the scope indicated by the logic model. A rough preliminary model is shown in figure 10-9. Note that the area beyond the double lines on the right has not been included in the area of interest presently bounded by the school system’s combined logic model. For purposes of this example, that right-hand space is no longer considered. If the evaluation had been commissioned at the state level, even the preliminary model would, of course, have been more extensive.

The functional model of figure 10-9 (with the demonstration shown as a black box) shows the believed inputs, contributions, and outcomes of the knowledge-transfer intervention, plus the comparison group (group Y). The logic has been used to further scope the testable functional model since inputs and contributions to the process have not included the noise from in-house activities such as the superintendent’s speeches to the National Education Association.
Figure 10-8. Laying Out a Direct Intervention

Figure 10-9. Black-Box Testable Functional Model of CAI Program
Figure 10-10 shows some of the elements that might appear as the testable functional model becomes more detailed—that is, the model that represents the combined beliefs of those in charge about the process and structure of the intervention.

Unlike the logic model, elements in the functional model represent flow, function, and relationship of structure. Adding to the model or expanding some sections in detail must be done systematically, like a blueprint set—that is, structure and relationships are preserved. This can be illustrated by showing a first level of expansion of the detail in the testable model for just the instruction box (see figure 10-11). The flows are preserved, and more detail has been presented about how the instruction function takes place for groups X and Y. More-detailed representations could be prepared as needed.

**Discussion and Validation**

After laying out the testable information, the evaluator must first validate it—that is, the information from each source is written up and, if possible, reduced to the model form in which it will ultimately be used. The documents are then returned to their respective sources who validate or correct them so that they accurately represent what was meant. This step should be taken in an atmosphere of mutual helpfulness and prefaced with a statement like: Here is our layout of what we think you told us. Could you look at it and tell us where we have misinterpreted you or made errors in representing your beliefs and expectations? Testable information and models should always be reviewed, corrected, and sent back by the person or group whose views they are to represent. During validation, each individual source should be allowed to concur in, add to, delete, correct, or change anything in the testable information previously provided by that validating source (but not, of course, allowed to change someone else’s testable information).

This validating step not only ensures that the evaluators have an accurate representation of what those in charge profess to believe, but also it avoids the unpleasant consequences that result from silly statements that people wish they had not made. Remember that this is no time to argue about what might or might not have been said. Testable information—including changes made during validation—is accepted uncritically. It is important, and a great time-saver, to have testable information validated before presenting even early comparisons of testable with equivalency information.

The testable logic and functional models are used as the bases for several very different discussions. The logic model is a preliminary vehicle for discussing semantics, identifying mavericks and scoping the area of interest. The semantic discussions are always interesting and
Figure 10-11. Additional Detail on the Instruction Testable Functional Model
sometimes surprising. As discussed in chapter 5, different levels of the organization often speak vastly different languages even though the words sound the same. The differences can seriously affect the logic and functional models and, ultimately, the evaluation design.

For instance, in one EA, the impact component, *then the rate of technological change will increase*, appeared in each of twenty logic models collected. However, when the participants were asked to define technology, twelve different definitions, ranging from a *widget* to the *industrial arts that govern production* emerged. The best time to discover and, if possible, resolve such semantic differences is during construction of the logic model. Failing that, differences can be resolved when testable and equivalency information are compared. In no event should the resolution of semantic differences be allowed to wait until after the evaluation has been designed. Without these early discussions the evaluators may find that they have spent a great deal of time and money designing a plan to measure something that is of little interest to either the person who must accept the plan or the person who can use the evaluation results.

The discovery of outriders can also have significant implications for both the evaluation and the program itself. An example can be drawn from an EA done for a regional rural organization that had been formed for the purpose of keeping land in the hands of small farmers. An outcome component of all of the logic models collected was then small farms will prosper. One model (that of a state director) however, contained an outcome component then people will be able to get food stamps. The visibility of that outrider component caused a change in the EA schedule. Equivalency information was quickly gathered in that state and it became clear that the state personnel had actually been diverted from legal and technical assistance aimed at keeping land in the hands of small farmers in order to help process food-stamp applications. The remainder of the EA was therefore suspended in that state until those in charge brought the outriders back into the mainstream and established a vigorous program monitoring system there.

The final use of the logic model is to scope the project. Figure 10-9 is a testable functional model of the CAI demonstration, and it delimits the investigation as agreed upon in discussions of the logic model. Note that the EA and subsequent evaluation is limited to the current demonstration and that the principal s impact component, We will hire fewer math teachers and more music teachers, is not included. Those scoping decisions should be made early, before resources are diluted, in an effort to find information that is not of timely importance. In this hypothetical example, the investigators might well have spent quite a lot of time collecting information about music and math teachers before discovering that the assistant superintendent was not the least bit interested.
The discussions based on the logic model are conceptual. The discussions based on the functional model are structural and analytical. The testable functional model serves as a framework for representing what is believed to be going on and for discussion of potential measurements and comparisons that those in charge believe can be taken in the direct-intervention domain and will constitute an acceptable level of proof of the efficacy of the program. The CAI program described is a very simple intervention and one that sits squarely in cell A of figure 6-3. The people involved know what the program is supposed to accomplish, are certain that it will produce the expected outcome, and know what significant others will accept as measures and comparisons. A simple computational strategy is called for. (A review of chapter 6 should reinforce the reader’s belief that most programs do not have it that easy.)

Figure 10-12 shows some of the measurements and comparisons that might result from discussions of the testable functional model. These are the measurements and comparisons that those in charge would like taken. As equivalency information is collected and interlarded with the testable information, the model and suggested measurements will gradually be amended to more closely resemble reality (more of this in chapter 13).

Warning

It is almost irresistibly tempting to spend most (or all) of the money allocated for an EA on gathering and modeling testable information. This is particularly true when articulate and thoughtful people are the major source of the information. Remember that all great works of fiction are created by articulate, thoughtful people. The degree of realism injected into such works can be assessed only by readers who are familiar with the described environment.

We recently saw a group of competent analysts spend nearly six months analyzing testable material about a portion of the Occupational Safety and Health Administration (OSHA) that consisted only of about a dozen staff spending less than $17 million a year. In the same length of time, these analysts could have examined almost every actual activity that the group under study had carried out over a number of years. In sum, the equivalency investigation would have been much simpler than the testable.

This is why we always suggest an iterative development of testable and equivalency information. One type of information serves to Scope the amount of effort that is to be devoted to the other. If you have one day, one week, one month, or one year to expend, it is wise to systematically divide the time between collecting and modeling both testable and equivalency information so that early comparisons between testable and equivalency models can be made.
Figure 10-12. Some Suggested Measurements Based on Testable Functional Model of CAI Program
Note

This Page Intentionally Left Blank
Equivalency Information

Equivalency information (and the two models of it) represents the activity being studied as its actual operation is observed by the evaluators. The two types of equivalency models—functional and summary statement—are diagrams of the unifying concept(s) and along those flows) as they exist in the direct-intervention process. The boundaries and functions of interest in the equivalency models are delimited by the testable information. Unlike the testable models, though, the evaluators are responsible for the accuracy of the equivalency models. Taken together, the two models of equivalency information lay out what is actually happening in the direct-intervention domain and provide the framework for discussions about which expectations are realistic and which measurements are possible.

The Preliminaries

Before collecting equivalency information, two preliminary steps are required and a third step is desirable. First, at least a rudimentary testable logic model must have been drawn in order to scope the investigation. In the last chapter’s CAI example, for instance, the assistant superintendent’s early logic model limited the EA to CAI and kept the evaluators out of the gym.

Second, a unifying flow(s) must be selected. At this early stage, the chosen flow will not, in all likelihood, be the final choice (see chapter 3), but it will probably not be far off the mark and will keep the field team from chasing will-o’-the-wisps. In the CAI example, the early selection of knowledge transfer would serve to keep the field team from an irrelevant investigation of, say, the development of social skills. Had later discussions disclosed that social skills were, in fact, of interest, the field investigation could easily have been expanded. The point is that it is not likely that something designated early as a unifying flow will later turn out to be irrelevant. It may not be as good a central organizer as some flow later observed in the field, but it will almost certainly be of some interest.

To illustrate, the Institute for Computer Sciences and Technology (ICST) at the National Bureau of Standards is currently undergoing an EA. In the course of the EA, equivalency functional models of several federal ADP installations are being constructed. At first, we thought that the life cycle of a computer would serve as a unifying flow. For reasons that we discuss shortly, this turned out not to be the case (we eventually used the transformation of data into information). However, the initial choice was close enough so that most of the equivalency information
gathered around that flow was relevant. Also—just as important—the early choice prevented our field teams from drawing equivalency models of the operating systems of typical computers—a resource-consuming task that they would almost certainly have engaged in.

Third, it is desirable, although not necessary, to have a preliminary testable functional model (even if only black boxes) before the field teams begin their work. This model is useful at this early stage both because it can be of help in selecting a robust unifying flow and because it will help the field team schedule interviews and select suitable background material. In the CAI example, for instance, the early knowledge that contributions to the CAI demonstration were believed to include CAI hardware and software, curriculum revision, and so on and that students were to be trained (see figure 10-9) would imply that team members should read up on those things so that they do not make fools of themselves in interviews and furnish categories of relevant people to interview while waiting to get on the assistant superintendent’s calendar.

In sum, if the evaluators should walk into a direct intervention naked of all knowledge of it, they would be in a situation analogous to watching a complicated game such as chess with no knowledge of the game’s purpose or rules. Certainly at first, probably for a long time, and possibly forever, it would seem as if the players were randomly moving pieces on a board. When watching a complicated game, it is much less puzzling if you read about or have someone explain the essentials of the game before you begin kibitzing a match. The early testable information—displayed as a logic model—serves the function of providing that initial briefing. The functional models identify potential flows and functions, the pieces on the board.

Building the Equivalency Models

In chapter 2 we discussed feedback loops as the operate in a mechanistic home heating system. In that discussion, it was noted that the system could be set up to operate open loop (no feedback) by, for instance, setting the furnace to burn a given amount of oil every hour and allowing, perhaps, for seasonal variations. This would, of course, occasion some discomfort among the household’s inhabitants during the not uncommon spells of unseasonable weather. It is also possible to operate an EA process open loop; but we do not advise it.

More is said later in this chapter and in chapter 13 on the subject of feedback while constructing equivalency models. For now, it is sufficient to emphasize that while simple intelligibility occasionally constrains us to discuss the creation of these models as if they flowed unimpeded from the pen, much backing and filling occurs as testable and equivalency information are brought together and compared throughout the progress of an EA. With that caveat, then, once in the direct-intervention domain, the evaluators try to trace the unifying flow(s) through it, recording those
points where something happens to it. The team should look for all the elements identified in the testable functional model, recording if, in fact, they appear in the direct-intervention domain and if so, where they are in relation to one another and to the unifying flow(s). A preliminary equivalency functional model can now be constructed, and from it, a summary statement can be extrapolated. (If a comparison or control group is involved, an equivalency functional model should also be constructed for it.)

The equivalency summary statement is now compared to the testable logic model. In some few (but memorable) cases, the differences between them will be so gross that the need for immediate action is evident. The small example of the putative tuberculosis-testing clinic (cited in chapter 9) or the large example of the Vietnam war are cases in point. When reality is sharply different from expectations, some big misunderstanding is usually involved.

The CAI example, which was adapted from an incident reported by former U.S. assistant commissioner of education, Leon Lessinger, to a 1977 Indo-U.S. seminar on educational technology, is an instructive case in point.¹

The school system’s request for funds actually came to the DoEd when it was the DHEW. It was received by the working bureaucracy in the Office of Education, and because of Dr. Lessinger’s known interest in CAI, it was called to his attention. Atypically for someone at the policy level of the those-in-charge domain, he took a personal interest in the direct intervention, reading the progress reports as they were submitted. He was particularly gratified by the comparison-group/experimental-group test scores that indicated that CAI was working very well. Because he had followed the program so closely, he had in his mind very complete logic and testable functional models.

After the project had been going for some time, Dr. Lessinger found himself in the vicinity of the direct intervention and stopped in at the school. He spoke to the principal, who identified the comparison and experimental groups for him and, at his request, brought him to the CAI laboratory. There, he found the expected number of carrels, each with a round hole in its side. It turned out that the holes were where the controls of the CAI devices would have been had the manufacturer ever delivered them. As it happened, the manufacturer had been unable to produce the devices and the program had never been implemented. What about the comparison-group/experimental-group test scores? The school system had committed what the seminar participants classified as a type V Error: outright lying. The true facts in the case were the following:

The school system discovered that the manufacturer could not deliver the hardware only after the experimental and comparison groups had been assigned.
The school system decided to develop a self-paced fifth-grade math course, using a textbook based on the material in the CAI course.

Group X attended the self-paced classes in a schoolroom where an accredited math teacher and a teacher’s aide were available to provide help whenever a student requested it.

Group Y attended a standard course with quizzes at the end of each lesson.

Had Dr. Lessinger been doing an EA, he would have drawn a rudimentary equivalency functional model and summary statement and gone back to those in charge to explain why, in terms of their present expectations, the program was unevaluable.

Figure 11-1 shows a hypothetical black-box equivalency functional model of the process. From that model, a rough summary statement can be extrapolated as follows:

If students use the new textbook based on the CAI course,

then they will be better educated than students who learn from the current textbook.

In a case of this type, the action taken might very well depend on who commissioned the EA. If those in charge were at the federal or state level, the most likely actions would be great moral outrage and immediate withdrawal of funds. If those in charge (and the funds) were at the school-district level, the actions might easily include a revision of the testable models to conform to the equivalency—that is, those in charge might decide to go on testing the new textbooks.
In that event, the EA team would continue gathering equivalency information and building models.

Assuming that to be the case the level of detail of the equivalency model would be expanded. Figures 11-2 and 11-3 show levels of detail similar to the testable models of figure 10-10 and 10-11 respectively.

It would be well to pause here and consider what might have happened had Dr. Lessinger not paid his visit to the school system but had, instead, commissioned an outcome evaluation based on the testable functional model (a not uncommon procedure). The outcome measurement and comparison suggested in figure 10-11 is the score achieved on a standardized math achievement test by the experimental group compared to the scores achieved on the same test by the comparison group. This comparison could have three possible results, each leading to a baseless conclusion and, possibly, to a groundless action. They are:

Result 1: No significant difference between group X and group Y.

Conclusion: CAI and traditional instruction are equally effective so choice should be the cheaper alternative.

Possible action: Continue traditional instruction and save the cost of the teacher’s aides.

Result 2: Experimental group does significantly better than comparison group.

Conclusion: CAI is significantly more effective than traditional instruction.

Possible Actions: (1) Continue experiment, testing the differences among classes led by teacher and teacher’s aide, teacher alone, aide alone; (2) order CAI equipment for the entire school system; or (3) Disseminate the results of the experiment to the nation of educators.

Result 3: Experimental group does significantly worse than comparison group.

Conclusion: Traditional instruction is significantly more effective than CAI.

Possible actions: (1) Continue traditional instruction, or (2) disseminate the results of the experiment to the nation of educators.

On occasion (as in the case of CAI), an EA will uncover gross discrepancies between the logic model and an early summary statement. These cases are always fun, but they do not happen very often. More usually, the EA team will gather a much greater quantity of equivalency information and develop more detailed functional models before significant differences, if they exist, become evident. In some cases (like
Figure 11-2. Some Elements of an Equivalency Functional Model of the Textbook-Demonstration-Program Direct Intervention
the regional organization mentioned in chapter 10), some number of field projects will be operating as expected while others will be delivering a surprise a minute. In other cases (like the child-health program in chapter 9), everyone will be doing something akin to the original design but no one will be operating exactly as expected.

In equivalency work, the questions of what to examine, in what detail, and how to describe structure and interrelationships require a robust unifying flow(s). The selection of an appropriate unifier is largely a matter of trial and error that is facilitated by iteratively examining the testable questions suggested by the models of those in charge and testing them against equivalency information drawn from direct observation. As the functional models are expanded and refined, more questions will occur.

Using a few key expectations and questions from the testable information, the EA team should attempt to locate, on the early equivalency functional model(s), those places where measurements must be made if the questions are to be answered. If measurements for key questions cannot be located or related to each other, a different unifying flow(s) should be tried.

As mentioned earlier, the life cycle of a computer was the first unifying flow tried during the EA of ICST. We found, however, that while the equivalency functional models drawn around that flow displayed enough information to locate measurement points to answer many questions of interest to those in charge, no significant information about end users appeared. By substituting the flow, data-to-information, we were able to locate measurement points for all of the relevant activities while still filtering out noise.

In some instances, more than one unifying flow may be required. We mentioned earlier that in the CAI example, the development of social skills might have been of interest to those in charge—particularly if guidance counselors were involved in early discussions. It is not likely that an equivalency model based on knowledge transfer would capture many socialization factors or vice versa. In such cases, the use of more than one flow is necessary (student flow might have worked well as a central organizer).

Occasionally, the EA team might try every conceivable unifying flow and still be unable to find locations on the equivalency functional model where answers to significant questions might be located. Should that be the case, then the program as it stands is unevaluable in respect to those questions, and those in charge should be apprised of the fact as soon as it is established.

Alarums and Excursions

Chapter 13 discusses the formal process comparisons and decisions about sequential purchase of information and the meetings in which the purchase decisions are made. Apart from those meetings, there are
usually many other times when the EA team gets together with some of
those in charge. Some of these ac hoc meetings are occasioned by the
discovery of something unexpected in the direct intervention. Note,
however, that these meetings are not for the purpose of confrontation. (It
is bad form to burst into the room crying, “Aha, you dummies, you don’t
know what’s going on in your own program!”)

Sometimes the EA team will discover that the unexpected find was
due to an incomplete testable model. This might happen for a variety of
reasons including (1) an important stakeholder had been overlooked and,
therefore, testable information not collected from that source or (2) a
testable source failed to provide complete information. In either event,
the information should be collected and testable models redrawn.

In those cases in which the unexpected find is unexpected by both the
EA team and those in charge, decisions must be made—namely, to
change the belief structure, to change the reality, or to call off the EA.
When the decision is made to change the belief structure, the testable
models are emended to reflect the new beliefs. At no time is the
equivalency model changed to conform to the testable model unless there
is a corresponding, observable change in the direct intervention itself.
The recognition of differences between the structure of the testable and
functional models and the accurate, valid determination of causality is
what EA is all about. Where differences exist between the models, the
evaluators should verify their own observations and comparisons. If they
still do not see what is on the testable functional model, they do not put it
on the equivalency model. If reexamination shows that they overlooked
something the first time, then the equivalency functional model is revised
accordingly. The point is that things are in the equivalency models
because the evaluators saw them, not because someone told them about
them.

Level of Detail

Some parts of the functional model will be drawn in great detail, other
parts will remain in a fairly rudimentary state. The development of detail
will depend on the loci of interest by those in charge and on what is
required to see if expectations are being met in ways that can produce
measurable results and causal analysis. The equivalency model is the
basic tool underlying causal analysis in any complicated situation. (Too
many failures over too many years with other methods argue for complete
descriptions of structural relationships as a guide to analysis as opposed
to analysts’ editing the descriptions to fit a particular analysis technique.)

In the CAI example, had the school system decided to go ahead and
test the new self-paced textbooks, those in charge would almost certainly
have been interested in the minutiae of the classroom instruction; it is
unlikely that they would continue to implicitly trust those people who had
lied to them before.
In the case of the regional rural group mentioned in chapter 10, those in charge were particularly interested in membership-recruitment procedures so that portion of the equivalency functional model was developed in great detail. A new youth program, which was more or less a sideline of the organization, was not developed much beyond the equivalency black-box model. Just enough was done there to assure those in charge that some activity, aimed in the expected direction, was taking place.

Similarly, those in charge of ICST are particularly interested in the process of developing standards and in the effects of usage on ADP centers in the federal government. Those aspects of the model are receiving focused attention.

In much of Nay’s regulation work, detailed case examinations drawn from particular examples in the field have often cut through years of speculation and prospective analyses about the effects of regulation.

**Convergence**

As equivalency information is gathered and modeled, it is compared with the testable information. Over the course of the EA, the testable models, representing the belief structure, are changed as the beliefs of those in charge are adjusted to resemble reality. Ultimately, the two families of models converge to form the evaluable model—a functional diagram that is an accurate representation (equivalency) of those parts of reality that are of interest to those in charge, that are needed to test their expectations, or that are required to determine the cost of answering their questions.

If all has gone well, by the time the evaluable model is drawn, the belief structure of those in charge (represented in the testable models) will resemble the reality of the direct intervention (represented in the equivalency models). If there are differences, they will likely be in the level of detail; the equivalency functional model normally shows sufficient detail to permit a wide panoply of potential measurements (and their associated costs) to be shown.

In the remaining sections of this chapter, we review some of the problems that might prevent the orderly convergence of the two families of models and then discuss a sometime thing: the hybrid testable-equivalency model. In chapter 12, we present some formats for displaying the evaluable model.

An analysis of the CAI example points up three situations that inhibit the convergence of the testable and equivalency models. They are

1. There is nothing in the direct intervention to measure.
2. There are many things in the direct intervention to measure, but none of them appears on the testable models.
3. The direct intervention is not well enough developed to permit the measurement of some crucial items of interest.
We now take up the three situations individually, using the CAI project as a common example.

**Nothing Measurable in the Direct Intervention**

In the CAI example, our hypothetical evaluability assessors, armed with testable logic and black-box equivalency models, would have discovered very early that there was little similarity between the beliefs of those in charge and the reality of the activity. The only thing in the testable models that had a counterpart in the direct intervention was the selection of the comparison and experimental groups. Based on the comparison between the two families of models, the project is clearly unevaluable. If the comparison reveals as much information as those in charge care about, about, the evaluation activities (and probably the project) stop right there.

**Many Things Measurable in the Direct Intervention but Nothing that Is Indicated on the Testable Model**

It may be that something is occurring in the direct-intervention domain but not what those in charge expected. This was the case in both the CAI example, the tuberculosis-testing project, and the Vietnam war. Suppose now that our hypothetical CAI evaluability assessors had returned to those in charge and explained that the direct intervenors had substituted teacher-assisted self-paced instruction for CAI. Those in Charge would have several options available: (1) They could abort the evaluation and the program at that point. (2) They could abort the evaluation but let the project run its wayward course. This might be the preferred option if the project was either so close to completion that aborting it would be meaningless or if the project was politically untouchable. (3) They could hire another team of evaluators to do an intensive impact evaluation based only on the testable model. This might be the preferred option if either bureaucratic superiors or a congressional-oversight committee was demanding an impact evaluation for the 1 percent set aside for evaluation. Potentially, this option is the most destructive since far-reaching policy decisions might be based on the misleading data produced (4) They could change what is happening in the direct intervention to conform to the testable model—for example, hold up funds until that CAI devices arrive and then proceed according to plan. To succeed, the implementation of this option must usually be accompanied by assiduous project monitoring to keep the project from wandering off course again. (5) They could change their beliefs about the program to conform with what is actually occurring in the direct intervention and proceed with the EA and the evaluation. This last option might be selected if it were politically expedient to proceed with the project and/or if the outcome of the intervention as implemented was of
of interest in its own right (both of these conditions might have obtained, for instance, in the tuberculosis-testing project). In this case, a new family of testable models (based on the revised expectations of those in charge), would be constructed and reconciled with the equivalency models to produce an evaluable model.

**The Direct Intervention not Sufficiently Developed**

There are a number of reasons why the direct intervention might not be sufficiently developed to enable the construction of a comprehensive equivalency model. In some instances, some, but not all, items of interest in the testable models may have been implemented. In other instances, the project may not have been running long enough to have reached the point at which significant measurements can be made.

**The Hybrid Testable-Equivalency Model**

In all of these situations, the testable and equivalency models will not smoothly merge to form an evaluable model—that is, a model that shows where in the observed direct intervention the significant measurements of interest can be made. In such situations, it will be necessary to construct a hybrid model including, wherever possible, representations of observed activities but filling in the gaps with described activities from the testable family.

In the CAI example, for instance, suppose that those in charge bought the direct intervention almost as implemented but also had some questions that could not be answered by the existing project. As an example, their testable models might indicate expectations that another experimental group, W, would be tested with only a teacher (no teacher’s aide) in charge. To that extent, the school system would be instructed to modify the project, and a hybrid model would be drawn. The measurements shown in this model would be based both on the equivalency model as far as it went and on fillers from the testable model where reality gaps existed. In constructing these hybrid models, it is important to indicate which parts are based on observation and which are based on expectation. As the intervention proceeds, close project monitoring will permit the gradual substitution of observation-based information for that grounded only on expectations.

Sometimes, the inability to construct a complete equivalency model may just be a result of inadequate running time of the project. If you look back at figure 1-1-3, you will note that the equivalency model ends before the semester is over and before the students are given their final exams. The equivalency model was abbreviated simply because the semester was
not over and the students had not been given their exams. Again, a hybrid model, indicating which parts were based on observed reality and which on expectations, would be drawn.

Remember, though, that these hybrids are preliminary. Reality may not develop as expected. At the extreme, for instance, some natural disaster may prevent the students from ever being given their final exams. Somewhat more likely, group X and group Y could be given noncomparable exams. When hybrids are used, close project monitoring is a necessity. This reduces the probability that something untoward will happen, and if, despite best efforts, something does go awry, both the evaluators and those in charge will be adequately forewarned. In either event, an intensive evaluation based on myth will be avoided. Figure 11-4 shows a hybrid model for a CAI project in which those in charge accepted the lack of CAI hardware and directed that a second experimental group, W, dispensing with the teacher’s aide, be added to the design. Notice that those elements that were not observed are indicated by broken lines.

Whether you are dealing with a true evaluable model based on equivalency information or a hybrid model, it is essential that the model be used as a referent for discussing proposed measurements, schedules, and costs. The ubiquity of the model, showing structure and function, is an invaluable aid in focusing attention on the doable, the desirable, and the economically feasible.
Note

This Page Intentionally Left Blank
When Is A Program Evaluable?

As noted in chapter 9, we originally developed EA in the early 1970s as a way of avoiding large, useless evaluation. Today, EA is more often used to test expectations against reality in government operations and to bring the expectations and reality of government into closer agreement, with or without the intention of eventually performing a large, formal evaluation.

This shift in usage perhaps grows out of our original premise that most unevaluable activities of government got that way because they were unmanaged. Guidance for the ship of state requires more than legislation and a senior-executive service. Reality-based management, inspired by reality-based congressional oversight, can do more to answer the question, What is government capable of doing?, than endless theoretical debates about what went wrong. Again and again over the last decade, programs whose testable models did not match equivalency have broken their bows on the reefs of reality. And, as we saw in the analysis of the CAI program, conclusions drawn from evaluation of unmanaged, unevaluable programs are baseless and, in some cases, downright dangerous.

As pressure for effective congressional oversight mounted, the comptroller general noted that:

GAO had developed its suggested procedure for congressional oversight as an alternative to Senator Leahy’s proposal for an evaluability assessment which attempts to determine whether a plausible program has been implemented before resources are committed to a costly evaluation effort (S.R. 307 introduced in the Ninety-Fourth Congress). After attempting to apply the assessment approach to selected pieces of legislation, GAO concluded that the question is not whether it can be done theoretically, but how it can be done in a way which will provide results useful to the Congress. Since that report was issued, GAO has continued to study experience with evaluability assessment. Such evaluability assessment is an important step in GAO’s suggested procedure for congressional oversight.\(^1\)

Given this movement to a much wider and more-active scope for EA, the questions today are focused much more on how to get the testable and equivalency models to converge to an evaluable form rather than with simply coming up with a finding of evaluable or unevaluable. This results in two consequences:
1. The EA process goes through a series of short iterations with lots of feedback to key stakeholders (see chapters 13 and 14).

2. The definition of evaluable needs more clarification than we had initially given it.

We had originally stated simply and clearly (we thought) that an evaluable model(s) is a functional diagram(s) that is an accurate representation of those parts of reality that are of interest to those in charge, that are needed to test expectations, and/or that are required to display the methods and/or costs of measurement and answering questions.

After observing applications over the last seven years, we occasionally feel like a parent at a PTA meeting. We sometimes have a strong urge to yell: What are you doing with my child? At other times, we have found the results to be immensely pleasing. In most questionable uses of EA, we find that steps or parts of the process that we originally thought were obvious were left out for one reason or another.

Part of the confusion stems from the fact that one early published article dealt predominantly with testable questions. We did not think to add the caution that policy expectations are only half of the process. As it turned out, such a caution was needed; some readers did not realize that comparisons with equivalency were also required. The early GAO tests on using EA in congressional oversight showed conclusively that both testable and equivalency information were necessary. This particular result did not surprise us but is, apparently, still unknown to the entire world.

In any case, a better definition of what it means to be evaluable, along with paradigms, formats, and worksheets for deciding, are presented here subject to future revision as we grow smarter. The material in this book is primarily directed toward federal-program operations. This book does not incorporate all that has been learned in the last three years of applications in regulation. By necessity, the examples are simple since complex investigations tend to generate yards of functional flow diagrams. For example, an equivalency diagram by Sam Anzelmo, which represents the Baltimore criminal-justice system from police custody through release from prison parole, took its creator about four months to complete and was twenty-two feet long.

**Evaluable Programs**

In most cases, a program is evaluable when it meets the criteria shown in table 12-1.

EA is a process illustrated in this book by table 12-1 and by figures 9-1, 12-1, 13-1, and 14-7 and their accompanying text. The EA process is entered into, continues iteratively, and has certain rules for leaving: (1) Stop if it is undoable, (2) stop if everyone is satisfied, or (3) stop if
monitoring and evaluation can complete the work.
When Is a Program Evaluable?

The first comparisons in an EA are those between the models (Are the structures and relationships those expected, believed, or reported?). Then come comparisons of the expectations and questions with the measures that can be taken and the analyses that can be performed. The results of these comparisons are presented to those in charge and key stakeholders who decide what to do next. Comparisons can be made both/either after short periods of work and/or after long periods of work. The work times must often be tailored to both the situations and the ability to muster the needed knowledge and layouts for a problem.

As table 12-1 and figure 12-1 make clear, judgments on the evaluability of the program are nested and built upon preceding judgments. Positive judgments require that successful comparisons be made between the existing operation and structure in the direct intervention and what is believed or expected.

The next step is to define measures acceptable to stakeholders (from testable information) that can be taken in reality (from equivalency information). This forces further comparisons and often shows where more detail needs to be gathered. The comparisons and analyses through which answers are to be produced must be agreed upon and tested to see if they are feasible to carry out. The analytic approach taken can often greatly alter the answers obtained from the same operation. The users for

<table>
<thead>
<tr>
<th>A Program Is Evaluable When It Has</th>
<th>Which (Who) Are (Is)</th>
<th>EA Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure and operational relationships</td>
<td>Defined and in place</td>
<td>Equivalency functional model of the direct intervention is developed.</td>
</tr>
<tr>
<td>Agreed-upon key expectations</td>
<td>Feasible and attributable to the direct intervention</td>
<td>Testable expectations are located on functional model.</td>
</tr>
<tr>
<td>Agreed-upon potential measurements</td>
<td>Feasible to take</td>
<td>Measurement points and measurements are located on an equivalency functional model.</td>
</tr>
<tr>
<td>Defined, agreed-upon potential comparisons</td>
<td>Feasible to make</td>
<td>Structure of operation and nature of measurements are developed.</td>
</tr>
<tr>
<td>Uses for the information to be purchased</td>
<td>Capable of acting or effectively recommending action</td>
<td>Action options are developed.</td>
</tr>
<tr>
<td>Value to the users of knowing various evaluation outcomes</td>
<td>Far in excess of the costs of conducting monitoring and evaluation</td>
<td>Sequential purchase is used to gain incremental insights into overall costs.</td>
</tr>
<tr>
<td>Links to the direct intervention through which action based upon monitoring or evaluation information will come</td>
<td>Described, plausible, and in existence</td>
<td>The oversight and those-in-change structures are laid out in sufficient detail to permit confident action.</td>
</tr>
</tbody>
</table>
Figure 12-1. What is Evaluable? Two Kinds of Comparisons
the information must be located and their options for further action explicated. The value of the information to be gathered is assessed in terms of changes that the new information might foster (expected utility). Sequential purchase can often be used to clear up, in steps, how expensive information may be to obtain, what results are most and least likely to occur, and what is the value of knowing the answer in terms of the value of future actions. The value should seldom be a small and marginal number. It is best to attack situations in which the value of knowing greatly exceeds the cost of finding out.

We have included as a criterion for evaluable (because of the use of estimated utility) the links through which those in charge are to act. In some cases, analysts will go ahead without the last condition, expecting that if the information can be created the results will be so obvious, so valuable, or so startling that mechanisms to use it will be rapidly created. Sometimes this works and sometimes it leads simply to frustration. For instance, at GAO, and in departmental program evaluation, the presence or absence of links often determines whether the work is a cooperative program audit (links in place to use the information and a desire to know the results) or a so-called black-hat audit (program will not cooperate, no management method available to use the results, management does not want to know the answer).

Either approach (cooperative or black hat) may be appropriate, and in practice they may often fade into one another over time. In the cases of the Vietnam war, New York’s financial situation, and the current economic situation, black-hat studies were required to bring the true situations to the attention of those in charge who did not want to hear. Cooperative arrangements were then required to make progress.

Even in cooperative studies there are many turning points after comparisons where tough factual positions must be presented and consideration insisted upon if the work is to succeed. We do not see a side to take in the arguments about cooperative versus black-hat studies. The same underlying work must be done to do either one correctly, and either stance may be appropriate in different situations or, in fact, at different times in the same situation. We would not always stop the EA if the last condition of table 12-1 is not met, but we have included it as an important consideration.

Comparisons and Tests of Evaluability

This section contains a format for using evaluable models (or, in a pinch, hybrid models) to test projects for their evaluable and to help anchor discussions in reality. It is important to realize that use of this format is no guarantee against sloppy analysis or mistaken assumptions. It is simply a convenient way of displaying information so that the concerned parties have a common framework for discussions.

Several examples have been used to illustrate the format. The first ex-
-ample is the hypothetical EA of the CAI program discussed in earlier chapters. That example is followed by a discussion of typical errors that can creep into analysis of this type. Subsequent examples have been drawn from Nay’s files of actual EAs. These examples are not accompanied by error discussions. (However, if you should uncover some mistakes, we will claim that we knew about them all along and were just testing you.)

Figure 12-2 shows the basic format for displaying the information. The top block is for the evaluable model, on which measurement points are located. The middle block is for the potential measurements ($M_1$ through $M_N$) that are linked to the measurement points by broken lines. The bottom block displays the potential comparisons and their relationship to the questions and expectations (question 1 through question $N$) of those in charge. A (●) in any column and row shows that a given measurement is used to address a given issue. The ⊗ symbol indicates that a comparison is to be made.

**Textbook (CAI) Demonstration**

In the CAI example, comparisons of testable with equivalency models pointed out something of major interest—namely, that no CAI was involved. To measure the actual experiment, either figure 11-2 or figure 11-3 could be used, since they reflect what is really being carried out. A hybrid model could also be used, if necessary, in order to include intended measurements of great importance to those in charge. Remember, though, that if a hybrid model is used, testable information must be identified and closely observed until it joins the equivalency ranks.

Figure 12-3 is a hybrid model (based on figure 11-2). A hybrid was selected as the evaluable model in this example because one measurement of great importance to those in charge—that is, comparative scores on standardized tests—depended on an event that had not yet occurred when the model was drawn. It was an intended event and thus, by definition, not equivalency information. When (and if) the event occurs, the broken lines will be made whole.

The hybrid model is the one shown in the upper block of figure 12-4. Potential measurements are arrayed in the middle block and linked to the measurement points (shown as small circles on the evaluable model). The right wing of the bottom block lists some of the questions of interest to those in charge. Comparisons are shown in the adjoining rows.

The first question, Are CAI scores better than teacher-paced instruction scores?, had been deemed unanswerable earlier when the first logic-model/summary-statement comparison was made. It is included in
When is a Program Evaluable?

Figure 12-4 only to illustrate a case for which measurement cannot be located on the evaluable model—a question that cannot be answered at any cost within the framework of the program.

The second question, Does the self-paced group advance further than the traditionally taught group?, might be answered by using math achievement scores administered pre- and post-treatment and compared within each group and between the groups, permitting the degree of achievement increase to be measured.

The third question, Is instruction time equivalent?, calls for a simple measurement and comparison of attendance in class.

The final question listed, Was the group assignment biased?, might be answered by a comparison of the preadministered match achievement scores of both groups.
Figure 12.3. Some Elements of a Hybrid Functional Model of the Textbook Demonstration
As we noted earlier, the use of a format, including this one, is no proof against error. The advantage of this format is that it surfaces measurement and comparison plans early and displays them in a manner that makes it difficult to hide mistaken assumptions and faulty reasoning. With that in mind, we can reexamine the questions and the proposed measurements and comparisons.

The measurements and comparison proposed for question 3 involved the use of preadministered match achievement tests. As the example was set up in earlier chapters, the evaluability assessors did not enter the case until after the quasi experiment had been running for some time (at least a few quiz results were in). Since there was no indication that the research design had included the preadministration of such tests (there is no box showing it on the model), the displayed plan for measurements and comparisons evidently assumes the existence of baseline data gathered from some other source. This blind faith in the availability and accuracy of baseline data is what we call *the economists' assumption* and is typical of analysts who are used to working with macromodels and relying on some outside agency for ex-
ample, the U.S. Census Bureau) for data. Should it turn out that pretreatment scores are not available, the only measurements and comparison left to relate to question 3 would involve the post-treatment scores of the two groups.

Question 3 itself contains what we call the what-do-you-mean-by-that? syndrome. The term instruction time is not defined and might or might not include time spent waiting for assistance, taking quizzes and—particularly in the case of the self-paced group—studying at home with the textbook. Failure to recognize and clarify definitional ambiguities early inevitably leads to much trouble with later analysis.

The last question, Was the group assignment biased?, calls into question the suitability of random assignment, as well it might. The application of randomization (or other statistical techniques) to small groups is a condition we call the statisticians’ ego trip. If we assume a student pool with a standard deviation of one relative to the mean score on a math achievement test and further assume that group X and group Y each consists of twelve students randomly assigned from that pool, there is about an 84 percent probability that the two groups are representative of the student pool and (by extension) are similar to each other. Note that the proposed method of checking the bias of the group assignment is a comparison of the preadministered math achievement scores. One might reasonably ask, then, why (if the scores are sufficient to be used for verification) the scores were not used as a basis for assignment in the first place? In other words, why settle for an 84 percent probability of similarity when close to 100 percent can be achieved? Statistical techniques were created to deal with large numbers; when small numbers are being considered, their use is questionable.

This brings us to the final category of error, which we call the educator’s fallacy. In this hypothetical example, the school system intended to make a systemwide decision vis-à-vis the purchase and distribution of textbooks based on one year’s test results from two small classes in a single elementary school. There was neither indication that it was known whether the school was representative of the system as a whole nor comfortable assurance that the classes were representative of the school. Strange things can happen in hypothetical examples. However, consider that the Office of Education actually funded a similar research design built around CAI equipment (much more expensive than textbooks), and further consider that the school system intended to make the systemwide decisions despite the fact that the devices never arrived. Even stranger things can happen in real life.

Child-Health Project

In chapter 9, equivalency models were shown for two child-health-project sites. Comparison of equivalency models with the testable model showed variations in operation and absence of health education,
rendering some desired expectations unevaluable. Questions must also be tested against each project equivalency in this case to see what is being measured before cumulating answers across all projects.

Figure 12-5 shows one of the projects laid out in the same format of model, measurement, comparison, and questions.

_The Clean Air Act and the Pennsylvania Electric Company (Penelec)_

The Clean Air Act Amendments of 1977 contained two sections that were aimed at stimulating technological innovation. Figure 12-6 presents the logic of these sections of the law. The Performance Development Institute’s Regulatory Center analyzed and followed up this section of the Clean Air Act. The logic extends over many domains, but we deal here with a direct intervention at a power plant that was submitted to the EPA in 1977 as a proposed technical-innovation waiver.

The logic shows the series of steps that was envisioned. In this example, Penelec was interested in a technological innovation that involved cleaning sulphur from coal before burning to produce electricity. This was proposed as a substitute for scrubbing the exhaust gas in the smokestack after the coal and sulphur had been burned [scrubbing removes the sulphur dioxide (SO₂) from the stack, preventing it from entering the air]. The proposed waiver was for a time period to try the innovation and to test it in practice with projected savings in cost, emission reductions over the three plants, and no production of scrubber effluent (a non-air pollutant).

The top of figure 12-7 shows a hybrid model of the proposed intervention submitted under the regulation. Two present generating units are required to meet a particular emission limit for SO₂, and a new unit is being built that must meet a much lower limit. Penelec proposed that instead of scrubbing out the pollutant at each stack, a large-scale, innovative, coal-cleaning technology would be installed to remove the sulphur from the fuel before burning (this intervention is shown in the bold box). The power plants burn the cleaned coal, generate electricity, and emit stack gas and a water effluent (not shown). In this case the measurements are needed over time, before and after the intervention.

Using the format suggested in figure 12-2 again, we can take a first cut at testing expectation questions against the model. Are emissions reduced or equivalent? The answer required measurement of SO₂ emissions and power generated over time, which are shown as coming from the plant and the stack. They can be compared over time and with the EPA limits. This question appears evaluable.

Are non-air-environmental impacts improved? This would require
measurement and comparisons of the scrubber and coal-cleaning-effluents. If it is to be made evaluable, more detail needs to be gathered on effluents and added to the diagram, along with methods of measuring them.

Are costs of innovation lower? Prices were obtained for the scrubbers being replaced (not on chart) and can be compared with the capital and operating costs of the coal-cleaning unit as it was built and operated. This question appears evaluable.

Does Penelec profit by the innovation? This would at first appear evaluable since costs and substitute costs have been identified, and if the coal-cleaning costs are lower (all other things being equal, as economists say), then cause and effect might be assumed. However, in this case there is an unfulfilled step in another domain. While all of this activity was taking place, step four of the logic—agency action to grant or deny the request—did not occur. The request was neither granted nor denied. At this writing action is continuing on this 1977 request. In the meantime, the plant and coal-cleaning equipment have been installed and are in operation. Thus, this question is not completely evaluable in using infor-
Figure 12-6. Logic Underlying Sections 111(c) and 113(d)(4) of the Clean Air Act Amendments.
mation available in the direct-intervention domain. Other information is required on potential additional regulatory or equipment costs, which could go as high as $100 million.

Does the innovation work? Measurements from this domain can be taken over time and compared to determine what actually happens technically. If work includes a completely satisfactory outcome, then other information about expectations and future regulatory actions is required for comparison.

The example shown here has been used to illustrate the kinds of questions opened up by the demands of table 12-1. The real case, detailed in the cited report is, of course, more complicated. The requirements of chapter 4 must be met for any measure to be used in practice. The methods of performing the analyses, sketchily indicated on the format, must be shown. Finally, information from other domains (not shown in the model) is required in order to render three of the questions evaluable.

The next example, again simplified, shows multiple domains.

*The Louisville Air-Pollution-Offset Bank*

Most of the material in this book emphasizes the direct intervention, using that domain to illustrate the key steps in the EA process. While Part III did discuss the existence and importance of the other domains, and the Penelec Homer City example illustrated how important (and how much trouble) those in charge can be, this air-pollution example is the first time we have added the those-in-charge domain to our flow diagrams.

This example covers a rather simple arrangement under the Clean Air Act for the banking of emission-offset credits. Under EPA regulations, each point-source pollution emitter is permitted to pollute the air to a given extent. If a company reduces its pollution an increment below what it is allowed, that increment of reduction can be placed in a bank and drawn out later when the company needs it for use, or the credit can be sold to another company that is building additional plant or increasing existing production or facility.

Figure 12-8 shows a rough flow model of the arrangement. Plants at time \(t_1\) have existing pollution and existing credits in the bank. When a firm is making a decision either to modify an existing plant or to build a new one, for example, part of the planning involves obtaining local and state approval of their emission permit. Part of the planned emissions can be covered by drawing previously banked credits of their own or by purchasing those of others.

Once the permit is issued and the new (or modified) emission source is in operation, the banking credits are adjusted. A similar process is involved when a plant takes steps to reduce emissions. These emission reductions can then be banked.
Figure 12-7. Hybrid Functional Model of Homer City Plants and the Direct Intervention: An Innovation in Coal Cleaning
Inserting this model and some sample measures and questions into the format of figure 12-2 produces figure 12-9. In this case, the measurements have been staggered to indicate the domain in which they are located. This also begins to illustrate a way to keep track of some practical problems that may need to be discussed during the EA regarding the location of measurement points.

For instance, what is to be accepted as an emission measurement? It could be measured in the local airshed, at the source, estimated from plant parameters, drawn from firm management reports, or lifted from permit data. Figure 12-9 indicates which is intended. This figure shows emissions being measured in the air near the plant.

This figure, unlike the previous three, introduces a situation in which the universes and measurements are not nearly so well defined (for example, as with a single CAI experiment, twenty-nine child-health projects, or a particular power-generation site). As broader policy models are drawn, it becomes necessary to do some careful measurement just to establish the appropriate universes. How many plants are in Jefferson County now? What are their emissions? What new plants and modifications have been built? What are their emissions? In these cases, iterative sequential work at balancing the universes, the cost of knowing, and what key stakeholders are willing to accept as surrogates can come into play quickly. the format also provides a compact, handy reference for discussion of these issues.

Summary

An evaluable program is one in which:

- The structural and operational relationships are defined and in place.
- Agreed-upon key expectations are plausible and plausibly attributable to the direct intervention.
- Agreed-upon measurements are feasible to take.
- Proposed comparisons are agreed upon and feasible to make.
- Users for the information can act or effectively recommend action.
- The value of knowing the evaluation outcomes far exceeds the cost of doing the work.
- Described, plausible links from those in charge to the direct intervention can effect change.
The last criterion is not absolutely vital to the evaluability of a program, but it is an important consideration in selecting a program to evaluate (similar to the consideration of great height in drafting an NBA basketball player).

An evaluable model is one that displays, in the form of a functional diagram, those elements of reality that are of interest to those in charge, that are needed to test expectations, and/or that are required to display the methods and/or costs of measurement and answering questions. The evaluable model is usually from the equivalency family. On occasion, however, a hybrid testable-equivalency model can be used. This is most often warranted when a planned, but not yet operating, function is expected to be added to the direct intervention. In those cases, the testable information must be clearly designated as such.

The second section of this chapter presented a convenient format for checking and displaying the evaluability of a program vis-à-vis the first four criteria. The format utilizes the evaluable model and helps to anchor planned measurements and comparisons in reality. The format is not dunce-proof; but neither is anything else.

Multiple domains as sources of data are discussed further in chapter 14. Sequential purchase of needed information during EA is treated next, in chapter 13.

Notes

4. Sam Anzelmo, Jr., Existing Adult Defendant/Offender Criminal Justice Process, City of Baltimore.
This Page Intentionally Left Blank
The Sequential Purchase of Information

Up to now, the coauthors of this book have collaborated in relative amity. In this chapter, however, we disagree about what is necessary to effect a smooth, sequential purchase of information. The disagreement is not fundamental to the process of sequential purchase: We are in complete accord that sequential decision making is the only sane way to conduct not only the EA but the program itself. The mixture of skills, talents, and training required to achieve that sanity is what is at issue.

The format of this chapter reflects our differences. The main body of the text was written by Nay and reflects his views. Where Kay differs, her arguments are stated in brackets. Where no bracketed text appears, either the relevant point has been made earlier or agreement should be assumed.

The transfer of EA to others has been found to work best through perhaps a year of on-the-job training as a team member under the direction of an experienced practitioner. No short course, training mechanism, instruction manual, or work description developed to date (including this book) has proved to be as reliable a method of transferring an understanding and facility with the EA process as hands-on training.

[Two points: First, a good short course, training mechanism, instruction manual, and several work descriptions developed to date (including this book) are infinitely more-reliable transfer agents than on-the-job training with experienced, incompetent EA practitioners (or experienced, incompetent doctors, lawyers, or plumbers, for that matter). Second, not everyone can get on-the-job training. Two EAs come to mind as illustrative. In one, I gave the team a short course in EA with emphasis on field procedures, and I participated only in the analysis; in the other case I simply supplied an early draft of this book, the field manual from the other EA, and heard no more about it until it was over. In both instances, the team made mistakes that I hope I would not have made. Also in both cases, however, those in charge were much better off than had they taken the other alternative allowed by available resources—that is, no EA at all.]

It is especially difficult to learn the sizing of, deciding on, and implementing of sequential purchase except through experience. Sequential purchase of information means that successive decision points are used to decide what part of the possible work—and how much of it—is to be purchased between one decision point and the next.

Actual experience may be necessary because the choices for sequential purchase are so rich and varied or because experienced
practitioners are much more skilled at deciding what to do than in
describing how it was done. Perhaps we do not do silly things as often
when we work as we do when we write. Most of the writing about
sequential purchase (including ours) is adapted from particular contexts
where it was developed (social-program evaluation, the National
Evaluation Program, DHHS health work, the ETIP Regulatory Processes
and Effects Project, and so on). Such adaptations therefore carry with
them (sometimes invisibly) attempts to deal with problems of that place,
context, and time that may or may not be part of the problem in a new
application.

For sequential purchase, the paradigm is figure 13-1. Testable and
equivalency information are being collected equally (review the section,
“Warning,” in chapter 10). Models are being sketched out and compared.

Figure 13-1. A Paradigm for Sequential Purchase
sis for the next decisions on sequential purchase.

The comparison step should be scheduled, distinct, and carried out as a specific effort. It should involve the charter under which the present work is proceeding, the work to date, a review group, comparisons (often requiring working meetings), presentations to those in charge, and decisions on the next step in sequential purchase.

[In our warning, we advised potential evaluability assessors to divide their resources equally between collecting and modeling both testable and equivalency information—if, that is, there are any resources left after all of the meetings are paid for. I completely agree with scheduling defined points at which to make and analyze comparisons. Who and how many are involved in the analysis depends in part on the resources available to the project. A good review team is expensive, and if you cannot afford good reviewers, you are better off without any.

A second important consideration is the attitude of those in charge toward meetings. Certain people love them—regard them as evidence that organized, disciplined minds are grappling heroically with the problem at hand. Other people regard them as committee meetings destined to produce the inevitable camel. When you are dealing with people like that, do not let those meetings show up in glorious profusion on your milestone charts. By all means, have the meetings you need, but keep them small, short, and out of sight.

The way you handle your meetings should depend on your judgment about the attitudes of those in charge toward them. Your ability to make that judgment also depends less on your experience with EA than on a combination of your innate good sense and your total experience (EA or any other) in working with those in charge.

I have been involved in two EAs in which the top level of those in charge served as a review team: They attended the meeting during which the first-cut analyses were debated, proffered review comments, and then later received the presentation wearing their those-in-charge hats. In both cases this was done in order to conserve resources; in both cases we explained the resource problem to those in charge and extracted their promise not to judge our competence on the basis of the first-cut analysis; and in both cases it worked out well. But I would not try it in all situations.

There are many ways to schedule, staff, and conduct the comparison steps, and the proper approach should be determined by the context. This is an instance for which standardization is a hazard.]

As long as the distinctions outlined in this part (chapters 9-14) are maintained, comparisons can be valuable at various times. The earliest comparisons that we use (sometimes called a quickset) involve gathering all available information (or people who have it) together and trying to lay out (over four hours to two days) the elements of an EA to see what is and is not available. A quickset involves some early comparisons between
what appears to be the state of testable and equivalency information and what obvious noncongruencies of logic or functional structure can be surfaced. We used to call this a logic comparison until we learned that comparisons are surer when performed between functional models and that there is no such thing as equivalency logic and thus substituted the extrapolated summary statement. This point is aptly described in chapter 9 in the section, “Testable and Equivalency Information,” especially at the end of that section. A quickset comparison may be expected to take some preparation, about a day to perform, and one to two weeks to document if written results are required in addition to the diagrams generated.

By contrast, we have seen the EA process used to continue to hold valuable comparison meetings over a period of several years after the effort commenced. These are usually cases of oversight or major program change and involve:

- Multiple powerful stakeholders whose interests are vitally affected by the outcome and changes to be made;
- Changes in testable expectations including policy, regulation, and law, either to match equivalency or intended to alter it;
- Changes in equivalency either to better match expectations or as a result of changed expectations.

One purpose of EA is to provide timely, relevant, accurate information that aims at program improvement through a modification of current operations. We consider these longer periods of EA and later comparisons appropriate (in oversight or evaluation) when real change and real convergence of testable and equivalency information are occurring.

When comparisons are presented to those in charge, potential decisions include:

- Stop. It is not worthwhile to go on.
- Proceed, but do not decide yet on the next major step.
- Facilitate closure between testable and equivalency functional models (and, of course, between expectation logic and equivalent summary statements).
- Stop. Everyone has as much information as is tolerable and wants the evaluation to end.
- Buy specific additional information (enlarging or altering the present charter).
Stop the EA because the program is now evaluable. Monitoring and/or evaluation can collect the remaining information needed.

We prefer a joint work team with members of those in charge and the evaluation team working together. The comparisons are initially performed by the EA team. Upon presentation of comparisons, those in charge will add or amend questions, fill in gaps in the logic model, revise functionals, and essentially attempt to resolve incongruities and inconsistencies in the work. They will usually retreat and/or attack if there are glaring problems in the comparisons of logic or structure. It is important to have validation of the testable working material from all key stakeholders before the comparisons are formed and presented. Keeping the times between comparisons shorter leads to fewer shocks and surprises. One of the advantages of EA is in working these problems through internally out of the glare of the spotlights, so that there are no embarrassing surprises before those in charge have a chance to act on the information produced.

The resolution of comparison findings is achieved through a process of negotiation—a process in which the evaluation team is an active participant. These negotiations are extremely sensitive, and any display of political clumsiness may totally destroy the mutual confidence on which a successful EA depends. Only the most experienced evaluators should participate in the negotiations with those in charge. [Two more points: First, experience in dealing with those in charge is unquestionably a requisite for participation in these negotiations, but the participants need not be confined to experienced evaluators. If your EA team includes neophyte evaluators whose previous experience included successful association with those in charge, they should be seriously considered for negotiation tasks. Second, not all experienced evaluators are good negotiators. A few fine field evaluators and many more fine analysts should be rendered mute and invisible when in the presence of anyone in charge.]

The working format displayed as figure 12-4 and the collections figure (figure 14-6) present the rich options for further investment. Almost as a sidelight one can see why the design of an evaluation or oversight scheme can be so difficult to carry out in a noniterative, nonsequential way. Additional work that may be needed at any point can include testable- or equivalency information collections, the models themselves, measurement design, selection or development of analyses techniques, question development or modification, and of course, EA to relate all of these things to each other as outlined in chapter 12. Additional domains may need to be examined. Information is obtained only at a price. Succeeding pieces of information, if well chosen, make the whole picture clearer. Following each comparison, the resources available are partially committed in a new charter that reaches only slightly beyond what can presently be seen.

When performing such work under contract, it is important to build
the mechanisms for sequential purchase into the contract. Few contract
offices or systems are constructed at present (in equivalency) in a way that leads to successfully processing a series of modifications in scope and approach as more is learned. The ability of those in charge to make sequential modifications should be a basic part of what is purchased under the contract.

[I have worked on joint teams both as an outside evaluability assessor and as a member of those in charge. From both perspectives this is a good arrangement. However, as one of those in charge, I have come to appreciate the difficulties of processing a contract that encompasses that type of close working relationship. If carelessly phrased, a joint team effort—particularly when coupled with sequential purchase—can be interpreted by federal contracts officers as falling under the regulations that govern personal-services contracts. That is an outcome to be avoided. Different federal agencies will react differently to the same language; go warily and check each agency you deal with for idiosyncracies.]

Some Examples of Why Costs and Effort Vary

Perhaps the reasons for sequential purchase can be seen a little more clearly if we consider important parts of each problem where varying amounts of money can be spent to investigate to various levels of detail or certainty. Only examples are shown, and no attempt is made to be exhaustive or complete.

Universes

In the child-health project, twenty-nine projects distributed nationwide were involved. At a high level of abstraction they might be the same, but operationally most all were different. Approximately 10,000 children a year were involved. These could be sampled or data about all of them included in treating various questions.

The CAI universe was small and manageable: two classrooms and approximately fifty children. That, of course, is what makes it possible to use it as an example without leaving many things out. The universe for Penelec’s Homer City Power Plants includes three power plants, local air, the immediate air shed, and a universe of air-quality monitors and of applicable emission-control equipment. Nearly any amount of money could be invested to study these universes depending upon the detail and certainty needed.

The air-emission-offset bank shows complexity in a local area: X plants, Y permits, Z offsets in the bank, plus several other potentially interesting or necessary universes (see chapter 12). The key is to buy only the information needed to address the questions correctly.

An interesting contrast between the model and the universes required can be seen from our work on hazardous-waste disposal. Figure 13-2 shows the simplest functional model of hazardous-waste flow, one reflected in the law, regulation, and supporting studies and reports. It
The Sequential Purchase of Information

shows only hazardous waste as a flow and that it is in one domain, and it reflects only generators, transporters, and disposers. The universes might be 750,000 generators; 10,000 transporters; and 30,000 disposers. However, suppose we begin to stratify and include more detail—affected by, for instance, 50 state regulating systems, 60,000 hazardous wastes, many methods of transport, and various means of disposal. Then the functional diagrams become more detailed (remember this step with child health and CAI in chapters 9, 10, and 11), and the universes appropriate for each case need to be further defined. Also, though costly, this is exactly what is necessary if much detailed measurement is really to take place; or indeed, if anyone is to find out what happened.

Domains

The multiple domains that occur in most real problems were treated in chapter 5 and are again in chapter 14 (they were touched on lightly in chapter 12). As an actual problem is laid out and the direct intervention characterized, suitable domains that must be considered in that particular problem emerge. The guides in this book are only an abstract way of getting started. Purchase of information from and about different domains (except the direct intervention) is a matter of judgment as the EA formats are assembled and linked together.

In child health, only the direct intervention with children has been shown in this work. Of course, the policymakers for the demonstration were another important domain that required a full mapping.

In CAI, the major domain was the demonstration itself although others were mentioned, including those in charge. In the power-plant case, only the plant domain was shown in the example. However, as reported there, action in another domain, EPA--technical-innovation-waiver approval or disapproval—proved to be of definite importance and was, of course, gathered.
The offset-bank example in chapter 12 displayed information in
domains including state regulation, local regulation, industrial-firm management, industrial plants, and the air. This begins to more realistically represent the complexity involved and show another source of variance of costs in sequential purchase. Each domain costs money and time to examine. In addition, the offset-bank diagram (figure 12-8) surfaces and displays additional unifying flows that were of interest.

Traces of Unifying Flows or Concepts

In the child-health example, we used children as the unifying flow. Not shown was the flow of health-care providers and from whence they came. Both were, of course, important.

For CAI, students worked well as one unifying flow. For the other we used a construct, or concept, that we have often used in education called knowledge flow. By this we mean where did the knowledge the children were to learn come from, what exactly was it, and who carried it? We usually try to include transfer to the child in a functional box. Years of observing education show that the expense of complete mapping in this area is highly variable. Consider, for example, social effects on the child (and how to find them out).

In the power-plant example, important traces of unifying flows included coal, exhaust gases, electric power, and effluent. We have alluded to, but not shown, the trace of another important unifying flow for this problem—the flow of Penelec’s waiver application. The preparation, submission, consideration, hearings, and discussions this waiver as the waiver travels through several domains are an important flow to trace in this problem and, of course, it was traced. The waiver flow tied together activities by many parties across many domains. Choosing to include the waiver flow and deciding how much detail to develop about this unifying flow obviously alters the amount of resources used or the proportion of the total available to be applied to this unifying flow.

The emission-offset bank in chapter 12 displays as traces the emissions, information, permits, applications, credits, and so forth. Also by employing the concept of time passage it traces industrial plants through time. Many points of potential measurement are displayed in figure 12-9. Each question considered can raise new potential measurements.

Measurements

Conducting actual measurement over a period of time is expensive unless one can piggyback on existing measurements. This means that each attempt to convert some characteristic of the world to a number can greatly affect the overall cost of the effort involved. We included in chapter 4 (table 4-1) some of the steps necessary to develop a
of them cannot be described, then the measurement actually has not been described and therefore cannot be costed accurately. The amount of validation of a measurement necessary in practice is a further cost. Our solution to this is sequential purchase during measurement definition.

What is not usually appreciated is the cost and effort that must go into constructing a measurement in order to guess whether taking it is worthwhile. The federal government has developed a facility for holding meetings and selecting long lists of measurements that everyone wants, which are then mandated for collection. The Paperwork Reduction Act of 1980 is, in part, an attempt by Congress and the OMB to stem this madness. [The Paperwork Reduction Act, like all pieces of legislation, is a rhetorical model. Only time and OMB will tell whether it stems the madness or just adds a new layer of it.]

Simply selecting measurements is obviously unacceptable in EA unless all of the measurements pass through the discipline of chapter 4 and chapter 12. Sequential purchase of information not only is essential here, but also it cannot work if the likely true cost of the measurement is not estimated to some degree; and estimating each one takes resources and time. [Spoken like a true contractor! Everything costs money and time, and those in charge cannot and should not divert all of the agencies’ resources to EA. Humbling thought though it may be, the most important work of any agency is producing its product, not evaluating it. As a rule, this means that the evaluation domain operates with insufficient resources to perform the ideal EA. Compromises have to be made, and some necessary steps have to be truncated or even eliminated. While you are doing all of this estimating, it would be helpful if you also estimated the costs of taking fewer and less-accurate measures. Those in charge would like to see an honest appraisal of the trade-offs. It is up to them to determine whether or not to take the risks associated with less-than-perfect information. The danger in not carefully spelling out the options and related risks is that those in charge may take an uninformed meat ax to the budget and inadvertently perform a prefrontal lobotomy when all along they had the option of just giving it a short haircut.]

We have found that even the most prolific suggester of measurements will usually return to reason if faced with the fact that some detailed equivalency work must be done in order to see if the measurement is possible and what it might cost.

In the child-health study, examples of measurement could range from children needing treatment to the hemoglobin level of a particular child. In CAI, the example measurements included attendance, instruction time, and math achievement scores. Passing the math-achievement-scores measurement through the steps of chapter 4, for instance, is a job for an expert and one that would have to be carried out, if the experimenter were not to be fooled about what has been measured. For instance, some tests use only differential questions (the ones that separate testees best), while others measure actual curriculum skills (multiplication). The whole
evaluation could go wrong on such a measurement point. Getting it right represents a variable cost until all the questions are known.

In the power-plant example, measurements included the sulphur in the coal and the $\mathrm{SO}_2$ emissions into the air. There are multiple ways to measure and points at which to measure each. Learning about these and selecting is a variable cost in the EA. Once again equivalency can often come to the rescue, and the Penelec power-plant case is a good example. If one is considering how to measure emissions, a trip to the plant and a day with the engineers may sound expensive. However, one can come back with a knowledge of what the measurement options are; how much the options cost to install, operate, and monitor; which are already in place and which must be purchased and installed; and what the operating peculiarities and reliabilities are. Alternative ways to do this are in common use.

The first alternative is to assign the task to someone to sit in an office and work out the answer. The second alternative is to have the measures selected by a committee of those in charge, to make the measurements, and to discover all these things during the application and subsequent analysis. Our observations over the last several years indicate that both of these alternatives are inordinately expensive and uncontrollable in both cost and achievement. Therefore, we recommend sequential purchase of the information during EA and inclusion of large doses of reality from field visits.

The fact is, none of these ways is as good as choosing an expert to do it. There are two corollaries: (1) The relative proportion of competent equivalency experts seems to be inversely proportional to the square of the distance from the application, and (2) a real expert, given this problem, will probably immediately make trips to some power plants. So it goes in health, education, industry, regulation, defense, and perhaps someday, economics.

In the air-pollution-offset-bank example, the number and size of actual banked credits and the amount of $\mathrm{SO}_2$ air emissions are examples of the measurements involved. A quick review of figure 12-9 shows the many possibilities that begin to occur when more parts of a program begin to be involved in an EA.

The format and the diagrams are merely shorthand reminders. Whether they are drawn or not, the program will still exist along with all of its measurement opportunities. Multiple stakeholders will still have questions and expectations to be addressed, and someone must keep track
<table>
<thead>
<tr>
<th>Elements</th>
<th>Child Health</th>
<th>CAI</th>
<th>Pensolc's Homer City Power Plants</th>
<th>Louisville Air-Pollution-Offset Bank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Important universes to be considered</td>
<td>29 projects, 10,000 children</td>
<td>2 classrooms, 50 children</td>
<td>3 power plants, Local air, Immediate airshed monitors, Emission-control equipment</td>
<td>Y Plants in Jefferson County, Y Permits for Operation, Z Offsets in Bank</td>
</tr>
<tr>
<td>Important domains of the problem</td>
<td>Health screening and treatment of children, Policymakers for the demonstration</td>
<td>The CAI demonstration</td>
<td>Generating plants, EPA</td>
<td>State, Local, Firm managements, Plants, Air</td>
</tr>
<tr>
<td>Important traces of unifying flows</td>
<td>Children, health care</td>
<td>Children, Knowledge Flow</td>
<td>Goal, Exhaust gases, Power, Effluent, (Waiver)</td>
<td>Emissions, information, money spent, staff, permits, applications, credits, construction, (plants through time)</td>
</tr>
<tr>
<td>Example of an important measurement</td>
<td>Hemoglobin, children needing treatment</td>
<td>Attendance, Instruction time, Math achievement scores</td>
<td>Coal sulphur, SO2</td>
<td>Number of Emission-Reduction banking credits, SO2 Air Emissions</td>
</tr>
</tbody>
</table>
of all these and tie them together.

As more information surfaces and is examined, more possibilities exist as well as more information for decisions about them. It would be a bold analyst indeed who would set out to meet the evaluable table (table 12-1) in any other than an iterative, sequential way. A summary of the examples from this discussion is given in table 13-1.

**Acceptance of Sequential Purchase**

Sequential purchase has a long, sound philosophical, mathematical, and practical history behind it. It was chosen for use in the EA process because both it and EA are aimed at practical attacks on large, important problems. Using it in government operations and oversight has been a long, tough battle. The budget process, the contracting process, the early thrust of social science in evaluation, and many other processes all claim surety rather than partial knowledge and do not at first seem to leave room for some of the concepts presented in this part of the book, although the concepts may seem valuable in everyday life.

Fortunately, the political process has many analogs to the EA approach and many uses for sequential purchase as well. Over the last several years, the comptroller general of the United States has succeeded in advancing these types of approaches both in his own work at the GAO and in uses for the Congress in legislative oversight. This may be partially because GAO also gets asked the answers to large, important problems and must produce answers that stand up under political, legal, and technical scrutiny and sometimes under stiff rebuttals from those examined. In the highest heat, the Chinese say, is where the toughest steel is forged.

Many of our colleagues have questions why we have written a book so technically simple. The intent of this book is not to engage in battles over what is science but to present these working concepts at a level that the person in government can grasp, examine, and accept or reject for use on the basis of the utility of the concepts. The concepts have been helpful both in daily work and in the strategy of governance, oversight, and evaluation.

Over the years of forming these concepts, we have used this material many times, and it has usually seemed to help and to bring results. We have found many changes that had to be made as we used it in practice. Since we are continuing to use it, the next revision begins with the publication of this book. [It may be correctly inferred from these bracketed comments that at least one of us began revisions before publication.] As it happens, we do believe that there is a long line of creditable science behind each of the chapters, perhaps more so than that behind several of the less-useful analytic devices that are often foisted off as complete solutions to the problems that this book attempts to address.
The authors are quite willing to engage individually or collectively in arguments with those who differ in debates, seminars, or reviews about either this approach or about science. We will continue to nettle people in this way as we have in the past.
Sources of Information for EA Models

The focus of this book—and of the modeling activities described in it—is the direct-intervention (or its surrogate) domain (see chapter 3, “Finding the Direct Intervention”). As we have noted throughout the book, and especially in Part III, the domain of those in charge (and even, occasionally, the evaluation domain) provides some crucial contributions to the direct-intervention process in the form of things such as legal authority, regulations, money, ancillary services, and so forth. Glitches in the flow of contributions (see chapter 9) can seriously affect the functioning of the direct intervention, and when this appears to be the case, the evaluability assessors should extend their modeling efforts to encompass the offending glitch and to trace its impact. Information gathered for these models might include both testable and equivalency information vis-à-vis the suspected “glitcher” as well as the glitcher’s testable logic and functional descriptions of the direct intervention.

In addition to collecting information from and about trouble spots, the EA team may want to collect information about the other levels of those in charge for a number of reasons—for example, to predict more accurately the measures that will be acceptable at various levels, to assess the relative importance of the various stakeholders, to judge the potential impact of a full field evaluation, and so forth. Figure 14-1 shows, generically, the levels at which information can be drawn and activities modeled. In any given EA, the generic labels would be replaced by those specific to the investigation (for example, authority equals congressional subcommittee on science and technology of the Federal Communications Commission; high rhetorical level equals president of the United States or secretary of the DHHS; bureaucratic leadership equals commissioner of primary and secondary education; and working bureaucracy equals program analyst).

While it is usually more difficult to gather equivalency information than testable information, once gathered it is not too difficult to recognize it for what it is—namely, something the evaluators have seen. Testable information is another matter. Descriptions, including laws and regulations that may have been written as much as a decade previously, come from all directions—from people waging vendettas against the people whose activities they are describing and from people in other parts of the organization (or even from the general population) who have not the faintest idea of how things work but who describe it anyway.
No EA can collect information from every person and about every activity that has anything to do with the direct intervention. However, particularly in large, well-funded EM, a tremendous amount of information can be gathered from key stakeholders who can affect the intervention, and the analysts thus will be hard put to keep it all straight without a checklist.

In this chapter we offer brief descriptions of the various types of information that might be gathered and suggest a format for checklisting the information once it has been obtained.

Sources of Information and the Related Models

Reality and Equivalency Models

The boxes at the far left-hand side of figure 14-2 shows how the generic levels of organization fall into the those-in-charge and direct-intervention domains. The first column of empty cells represents reality. The cells in that column are never checked; they are simply there as reminders that reality is the source of all equivalency information and is available as a reference when needed. Reality can never be wholly captured either by description or modeling; it is far too vast and complex. It exists, however (although it has a disconcerting propensity to change just when you think you have all the salient parts charted), for the evaluators to dip into again
and again (albeit at the expense of some time and money) as they make their equivalency models. As the evaluators develop the equivalency models, they will sometimes be confronted with apparent inconsistencies or dead-ends. When that happens, the evaluators should look again at reality because something important may have been left out of the model. The second column refers to the evaluators’ equivalency models of the activities. (It is important to note that these models are initially to be drawn using a pencil with a large and efficient eraser.) The information for these equivalency models is drawn from observation of the interconnected activities that occur at each level. For instance, the direct intervention includes the provider and recipient of service, and an equivalency functional model of that domain would show their activities, with the impact portion of the model extending on into the related general population.

<table>
<thead>
<tr>
<th>Location</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authority</td>
<td></td>
</tr>
<tr>
<td>High Rhetorical Level</td>
<td>Those in Charge</td>
</tr>
<tr>
<td>Bureaucratic Leadership</td>
<td></td>
</tr>
<tr>
<td>Working Bureaucracy</td>
<td></td>
</tr>
<tr>
<td>Service Provider</td>
<td>Direct Intervention</td>
</tr>
<tr>
<td>Service Recipient</td>
<td></td>
</tr>
<tr>
<td>Related General Population</td>
<td>Referent for Everything Else</td>
</tr>
</tbody>
</table>

**Figure 14-2. Sources of Operational-Activity Information and Related Equivalency Models**
When additional parts of the organization are to be modeled, one or more unifying flows are picked to trace through the organization (see chapter 3). At each level of interest, observations are made and discrete flow diagrams of the activities occurring there are drawn. The unifying flow that is selected to connect the vertical levels of the organization may or may not be the same as the flow used in any or all of the discrete-flow diagrams at each level. For instance, information and money may be used as the vertical flow (up and down through the organization) while people served may be used as the horizontal flow at the service-provider level. When equivalency information is gathered about a particular level or interface, the corresponding level or interface is checked on the chart.

**Testable Descriptions**

Rhetorical material is collected for four reasons:

1. To please the many important people who have produced it and thus want it collected;

2. To determine the beliefs of people who may be buyers of evaluation and users of results;

3. To ferret out expectations and questions that should be addressed in the EA and any subsequent evaluations;

4. To obtain beliefs about organizational and direct-intervention cause and effect in order to better focus equivalency-model work.

Figure 14-3 shows the same organizational levels as figure 14-2, but the two columns in figure 14-3 are for checking off testable descriptions of activities rather than equivalency information about the activities themselves. There is a considerable difference in the way the evaluator goes about getting the information to fill in the columns of figure 14-3.

In making their equivalency models, the evaluators observe identifiable operational activities (for example, counselor talks with applicant) and enter the operational activity on the flow diagram. The evaluators do not enter rhetorical descriptions on the equivalency model (for example, counselor advises applicant on realistic career opportunities involving a suitable career ladder consonant with the applicant’s educational and aspirational levels). In obtaining the information to be checked off in the equivalency-model column in figure 14-2, the analyst in a governmental organization will need all of the tricks from Part III and considerable on-site experience to pin down exactly what the activities (stripped of rhetoric) are.
In contrast to equivalency modeling, the testable descriptions checked off in the columns of figure 14-3 are shown in the models as being exactly what the people describing the activities say they are. For instance, if counselors report that they advise applicants on realistic career choices, that is the description represented by a check in the first column of figure 14-3 opposite service provider. If two counselors describe their activities in different ways, both descriptions are recorded.

The first column in figure 14-3 is reserved for checking off people’s
descriptions of their own activities and how they see their actions as interfacing with people at other levels. These descriptions can come from all levels, including the direct intervention. Warnings were given earlier not to be surprised if even the people conducting the direct intervention describe their own activities quite differently from the way the evaluation designer sees them occurring in the field. As in previous chapters, the use of the terms *rhetorical* and *testable* is not meant to imply value judgments on even the simplest of activities or rhetoric. Many times, simple activities are quite valuable in getting the actual work done. It is just that the governmental worker suspects (often correctly) that rhetorical descriptions are necessary to survive and will be quite hesitant to confirm simple descriptions of actual activities or to offer the activity devoid of its proper label. Thus, the person validating sick-leave requests and medical payments may be protecting the health of the workers. The busy school-system administrator whose time may actually be principally divided between personnel actions, school finance, and meetings with parents may be producing broadly educated good citizens. The federal-grant administrator who processes the grants for program Y may be changing the face of the United States through that program. The rhetoric may or may not prove to be true in the end. This rhetorical escalation of function is quite a natural phenomenon and one accentuated if grant-management teams, program-budget specialists, or other evaluators have visited the direct intervention previously.

Recently, another important factor has influenced bureaucratic self-descriptions—that is, the official position description (PD). Up until the Reagan reduction in force (RIF) in 1981, PDs were usually generalized descriptions of particular jobs written by unit supervisors and designed to permit maximum flexibility in hiring. They contained only as much specificity as was necessary to pass the often uninformed scrutiny of the Office of Personnel Management (OPM), the central personnel agency. They were rarely offered in interviews about the work being done.

The mechanics of the RIF changed things. When someone is “riffed,” that person’s organization is obliged to examine all of its PDs and, if a similar position at the same or a lower level is occupied by a person with less tenure, must offer that position to the “riffee,” who may then bump the current incumbent. When that occurs, the process is repeated until someone in the chain finally either voluntarily leaves the organization or has no position to bump into and involuntarily leaves the organization. At the inception of the RIF, some managers realized that PDs written to permit maximum hiring flexibility also permitted maximum opportunities for people to bump into the organization. Those managers immediately began rewriting the PDs (of people they wanted to keep) to be so specific that no one else could possibly qualify for the job. In many of those cases, the official PD provides a testable description that is not far from reality.

Within a few weeks, the personnel offices were flooded with newly rewritten PDs and, suspecting (in most cases, accurately) that this was a
mass attempt to circumvent civil-service rules, OPM declared a morator-
ium on rewriting PDs. This led to a wave of bumps by people who more or less fit the generalized PD but who in many cases were totally unqualified to perform the actual functions of their new jobs. In those instances, not only does the official PD not describe reality but also the reality itself is missing. This can be extremely puzzling to an analyst in search of a continuous equivalency model.

In either type of the present RIF-induced situation, the self-description furnished by the bureaucrat is likely to draw heavily on the written words of the PD. In either case, the motivation is self-protection. If, during an interview, the bureaucrat fetches out a PD and begins quoting from it, the analyst should be alert to the possibility that the self-description about to be unfolded will either be uncommonly specific and realistic or uncommonly general and fictitious.

Whatever the motivation for the self-descriptions, they should be accepted uncritically and logged in as part of the testable self-descriptive material. The important point is to distinguish this material, not only from equivalency information but also from other sources of testable information. The ability to find out rapidly who did what and said what about whom can save much later misunderstanding and days of analysis time.

Each level of the second column in figure 14-3 is for checking off the testable descriptions of that level in the organization as furnished by people at other levels. For instance, the service providers’ descriptions of what the working bureaucrats do is represented by the cell opposite working bureaucracy, and the bureaucratic leadership’s descriptions of how they believe the direct intervention operates are logged into the second column at the service-provider and/or recipient level(s). This second column thus represents accumulated descriptions of each level made by someone from another level.

As is probably evident from this description, the amount of information available (or easily generated) for logging in the testable-description cells can get out of hand. Interviews with all of the members of the DHHS or of your local school system, for instance, could prove both time consuming and boring.

There are at least three natural limitations on the number of interviews conducted (or material collected). First, the vertical flow selected through the organization (see chapter 7) will sketch a path though the activities of interest. Normally, only people along this (these) path(s) will be interviewed. Second, the natural good sense of the evaluators will demand a halt when the descriptions offered at any given level become redundant. Third, the time and money available for this stage of the evaluation may provide their own imperatives.

In limiting the interviews, however, keep in mind that it is often especially valuable to obtain rhetorical descriptions of the direct interventions from those levels that are expected to receive and act on the resulting evaluation information and from the level that has commissioned the evaluation design to begin with. Their testable
descriptions and expectations will often heavily condition the questions
they expect to have answered and the types of actions they may attempt to take when the answers are produced. These descriptions also serve to give the evaluators valuable clues as to what measurements and comparisons will ultimately be accepted by those in charge (see chapters 6 and 7). Some people interviewed will simply not have rhetorical pictures of other levels, but do not give up too easily in any specific case.

The products checked off during this exercise are two sets of testable descriptions. These sets of rhetorical descriptions form a significant portion of the information pool for use when drawing the testable models. Just as reality is the source of validation and further information in constructing the equivalency models, rhetoric is the source of testable models. The first testable set consists of people's descriptions of their own activities. The second set consists of collections of the descriptions of activities at each level given by people at other levels.

**Rhetoric and Writing**

Like Romeo and Juliet or Hero and Leander, bureaucrats and paper form a pair separable only by death. If one is to know the bureaucrat, it behooves the evaluator to seek out the paper. The first useful pieces of paper are the organization’s phone book, organization chart, program authorization, PDs, and so forth. These provide a rhetorical picture of the activities and actors and serve as the initial guide in selecting levels and people to interview.

Next, at any given organizational level, an identifiable stream of paper is emitted in an almost continuous flow. At the high rhetorical level, the paper is often in the form of records of congressional testimony, the reasoned analyses and recommendations of Cabinet-level committees, and exhortations to the people of the United States. Elsewhere there are progress reports, budget justifications, program descriptions, and even requests for proposals.

Apart from the paper that can be unambiguously attached to a particular level, a general rhetorical blanket usually covers the entire program. This is most typically found in the legislative history and the statute authorizing the program. Newspaper and magazine analyses of the program (particularly those articles written at its inception) often provide interesting testable sidelines. GAO audits usually prove helpful and often provide equivalency clues as well. While this general rhetoric is not often applicable to level-by-level testable models, it does serve a useful purpose in providing perspective and in piecing together a coherent logic model. Figure 14—4 is the checklist format for the written rhetoric.

**A Detour: Self-Descriptive Models**

Construction of testable functional models of the bureaucrats’ self-
descriptions is an intermediate step that may sometimes be necessary to make sense of the data logged in the cells of figures 14-3 and 14-4. It is often a good idea to sketch flow diagrams, relying solely on the testable descriptions provided by people of their own activities (leaving out other people’s view of each activity). Once that is done, the important elements of the self-descriptions can be compared with each other to see if they interface with the descriptions given by others (and found in the written rhetoric) of the same activities. This initial sorting and comparing provides a helpful preliminary tool for creating the model describing what those in charge think the activities are. It also helps in pointing out necessary additional interviews (or reinterviews).
The self-descriptive model will probably show a number of breaks in continuity where interfaces have been inadequately described or are described quite differently on each side. The self-descriptions of many of the actors will often be at variance with the description given of that level by others. These discrepancies and breaks in continuity can be found early. They highlight some of the points where the iterative interviews that produce the testable models should be concentrated. Many evaluation efforts go aground because the analysts are not aware that many disparate statements are made by people who are actually describing the same thing.

The final products of this step are flow diagrams of the actors’ self-descriptions and notes flagging important breaks and discrepancies.

**Testable Models**

The rhetorical source material about each level is the basis of a testable functional model in the same way that reality served as source material for the equivalency functional model. As each level of the model is drawn, its appropriate cell (figure 14-5) is checked off. In some cases, several checks will appear in a single cell, representing various unreconciled models of the same activity described by different people.

When testable models are completed, it is then possible to compare the testable models with the counterpart equivalency models in order to see how closely the two coincide. It was pointed out earlier that, if those in charge have reason to believe that important enabling and support objectives are not being met, they may decide on a preliminary evaluation at the suspect organizational level. Comparison between the testable and equivalency models is the stage at which the darkest (and truest) suspicions may begin to arise.

In making the testable model, there is an interesting contrast to the procedure used in producing the equivalency model. When the designers are in doubt about some section of the equivalency model, they return to the part of reality in question and attempt to determine its true nature more accurately. This is because the designers are trying to capture actual cause and effect and actual measurement points and measurement.

By contrast, when a fuzzy area is found in the testable model, the designers return to interviewing and either try to draw out clarifications or further definition. It is important neither to lose nor to abandon a fuzzy testable description. Even if made no clearer, such descriptions contain the essence of what people believe (or are willing to convey) about the organization and the program under study. Any evaluation to be designed and conducted must in the end be developed within this rhetorical context, or the evaluation is likely to fail. One reason for this is that for any answers developed by a full evaluation to be used, they must be interpretable by people who will still be operating within this existing rhetorical context when the evaluator returns with the completed
evaluation.
There is no attempt to alter the testable model independently to make it more realistic at this stage. In fact, all reinterviewing at this stage should be done under the flag of better definition, clarification, and confirmation that what was said or intended was captured properly.

This is especially important since, as the EA progresses, the designer will often have to ask those in charge to consider distinct differences that have been found between their testable model, other people’s testable models, and the equivalency model developed by the analyst. This can become especially touchy as a result of success measures’, heretofore undefined (within an agency), being defined in multiple-person meetings.

<table>
<thead>
<tr>
<th>Location</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authority</td>
<td></td>
</tr>
<tr>
<td>High Rehtorical Level</td>
<td>Those in Charge</td>
</tr>
<tr>
<td>Bureaucratic Leadership</td>
<td></td>
</tr>
<tr>
<td>Working Bureaucracy</td>
<td></td>
</tr>
<tr>
<td>Service Provider</td>
<td>Direct Intervention</td>
</tr>
<tr>
<td>Service Recipient</td>
<td></td>
</tr>
<tr>
<td>Related General Population</td>
<td></td>
</tr>
</tbody>
</table>

Figure 14-4. Much of the Rhetoric Is Written

Produced through Search of General Literature, Reports, Legislation, and so on

Written Descriptions of Activities Occurring at Each Organizational Level

Testable Descriptions of Activity from the Source of Authority through the Direct Intervention
Government Oversight and Evaluability Assessment

during the gathering of testable information. Occasionally, so much political capital may be expended in creating well-defined rhetorical measures for the testable model that later information about reality may become unacceptable—that is, it may be hard to get testable positions modified later if (when) reality is found to be different. This can be particularly troublesome if the testable expectations turn out not to match any of the possible measures that can be made from the real activities.

It is therefore quite important during the rhetorical interviews not to let those in charge lock themselves into positions that may have to be altered immediately after comparisons of testable and equivalency are begun.

Sometimes the testable models will reach from the authority through to the direct intervention; sometimes they will not. Often only a portion of a model or sequence of activities is found to exist after the rhetorical descriptions are reduced to testable-model form. These model fragments are often suspended in the air either in the those-in-charge domain or the direct-intervention domain, by misunderstood, or poorly developed, rhetorical linkages. Even partial models are often quite useful, however, in the final stages of EA. It is also not unusual to find multiple but conflicting models. In a 1969-1970 study of model cities, three distinctly different testable models were extracted from senior officials in DHUD, DHEW, and Office of Economic Opportunity (OEO). Even when all three were presented with the three quite different models, none chose to shift its views of what the program was about.

Ideally, though, the testable model should elicit affirmative answers to the following questions:

Is the linking of the activities of those in charge to each other, of those in charge to the direct intervention, of the direct intervention to the immediate outcome anticipated, and of those immediate outcomes to the expected impact on the overall problem laid out clearly enough to provide the basis for a rough logic model? A more-detailed functional model?

Have the activities of those in charge, the problems to be solved, the intended program interventions, anticipated immediate outcomes, and the projected extended impact been sufficiently well defined as to be measurable and testable?

The rhetorical descriptions and testable models represent the common context held by those from whom they have been taken. In part, the EA is to serve as a vehicle for bringing the questions that are asked and the answers that can possibly be obtained into better alignment. In some cases, either the testable beliefs or the actual activities (or both) will have
to be altered before further work can proceed. Contrasting the testable and equivalency models during meetings with those in charge as the assessment proceeds is an effective device for beginning this process of change. Mixing the testable and equivalency models together at an early point would lose the advantage of this contrast.

The beliefs held by those in charge at the time of the selection of an evaluation design may very well be the context into which the answers obtained during the execution of the evaluation will be placed and used. If the evaluation designer is serious about the use of the evaluation, mismatches between testable descriptions and equivalency models must be reduced or resolved either before the actual evaluation begins or during the conduct of the evaluation. The testable material (as finally agreed upon during discussions of the testable model) developed is one of the best indications of the positions people hold.

In figure 14-6, figures 14-2 through 14-5 are combined, representing the complete collection of sources, information, and types of models for the EA.

**Recapitulation and Coda**

The preceding material was spread over an unconscionable number of pages in order to impress upon the reader some subtle differences that can grow out of using multiple levels in the organization as information sources and both direct observation and rhetorical description as collection devices. Each information collection made (logged in a cell on figure 14-6) may expand (or contract) the need for another collection.

Sometimes the case is that the rhetorical descriptions of the direct intervention from the person commissioning the evaluation and the activities that are actually being conducted at the direct intervention are so unalike that any full evaluation will simply produce either the answer that nothing like that is happening or that a lot of information about what is happening is available, but nothing about the questions that were asked.

It is much cheaper, and much less painful, to discover that a situation like this exists during the EA rather than going to the field with a full evaluation (or a national disaster) to discover it. At this early point in design, those in charge can choose to alter their perceptions to be more like the program, or even to alter the program to be more like their perceptions, before a full evaluation is tried.

Figures 14-2 through 14-6 are really attempts to identify an exclusive and exhaustive set of possible information sources. The sources might be thought of as the indexes to a set of files that will be compiled during the data-collection phase of the EA. Should a full field evaluation be undertaken, the material gathered in and about the direct-intervention do-
Figure 14-6. Different Types of Information that May be Considered, Collected, or Created during EA.
main will contain the basis for subsequent design of an evaluation. Thus, the equivalency model is

The evaluator’s own best estimate, in measurable terms, of the activity and its actual linkages from the source of authority to, and including, the direct intervention;

A resource for determining how, or whether, important issues can be characterized in terms of measurable entities and what answers may be reasonably obtained through measurement;

A resource for determining the methods, potential costs versus accuracies, and the measurements and comparisons for answering specific questions;

The source for estimates of the value of partial or perfect knowledge;

The information that determines whether plausible links exist for taking action on results obtained.

In short, the equivalency model portrays the chains of effects that can be caused by decisions and actions at various levels as well as what information and interrelationships can be measured directly. The present nature of the intervention is one part of this model. In any subsequent full field evaluation, the direct-intervention portion of the equivalency model is usually used as the basis for defining the input, process, outcome, and impact measurements and their relationships to one another.

The testable model is the evaluator’s own best estimate, in measurable terms, of the various belief structures. It is

A catalog of expectations about the program,

The source of questions to which answers are requested,

A source for the believed accuracies needed,

A source for the value criteria for determining the believed value of obtaining various answers,

A guide to what types of information are likely to be acceptable and unacceptable at present,

A guide to the value of attaining different levels of precision and accuracy in the answers obtained through evaluation,

A set of cues for shaping the information created through evaluation so that it will be received easily by those in charge.

In short, the testable model portrays the symbolic communications of the organization arranged logically so that various parts of it can be tested and then either validated or rejected. This testing can be done by examin-
ing the logic of the agreed-upon attribution chains, measurement points, and measurements to see if they are plausible; by comparing the testable model to the equivalency model to see if grave or important differences exist that must be resolved before proceeding; or it could be done by taking the testable model as a basis for an evaluation and performing it. Of course, if the equivalency model differs sharply from the testable one, this latter course may be an expensive way of demonstrating the difference.

The equivalency model reduces reality to display the interrelationships and measurements that can be obtained from nature about questions of interest; the testable model reduces the rhetorical descriptions to display a set of testable measures and relationships agreed upon by those in charge as existing and underlying their questions of interest.
No one would expect to be able to analyze, direct, or fix a complex system without a good understanding of how it worked, what it did and was supposed to do, and how it was connected together. The would-be surgeon must know something of the complex human system—its arteries, nerves, muscles, and bones, along with their places and their connections.

Yet we often forget that government is a complex social system with many financial, political, informational, and social flows and feedback paths—the social-system equivalents of arteries, nerves, muscles, and bones. If you would not let a surgeon treat your problem without such maps of the body, perhaps you should demand them of your policy analyst or public reformer as well.

This book presents material that has been around for several years about how to develop such maps. An idea of how some of the material was developed is given in the appendix. Development has, of course, run far beyond this book. Several hundred studies or projects have been accomplished along these lines as the material developed. The authors have directed or been personally involved in over eighty projects, ranging in time from a few weeks to several years and in size from a single person to teams of large and varied size. As the preface mentioned, this book is just an introduction into the secrets that seem to make it all work.

The approaches are primarily for those who care to know as quickly as possible what is really going on and to actually take action to close the gap between desired or expected results and reality. We have found that the approaches are useful for management, oversight, and analyzing regulation, as well as for their original intent—that is, as a precursor of any formal program evaluation.

We have found that the evaluator’s bane so often reported in the literature—nonuse of the results—rarely occurs when EA is used. There is something about checking the expectations of important stakeholders against reality that stirs everyone involved to want to look over your shoulder as results come in and to take action to improve the comparisons before the results come out. Working iteratively and with sequential purchases of information encourages such interest and provides more opportunities for improvements to happen.

A detailed example that we regretfully deleted from the main body of the book makes an interesting point on the recurrence of problems such as the transmutation of government intent. Kay had prepared an example
based upon a federal demonstration in the communications area and had completely disguised it (in the interest of protecting the guilty) as a
program demonstrating garbage trucks. The example clearly showed the transmutation of government intent over many domains of action—congressional, executive, state, local, and so on—as the demonstration developed. When we circulated the draft to several federal officials for comments, one worried caller attacked the example as very thinly disguised and likely to cause several innocent people trouble. It turned out that those innocent people were responsible for a garbage-truck demonstration that had many of the same problems as the disguised communication example. We thus omitted the example. The point should remain, however: There are repetitive problems and patterns in government activity that must be dealt with.

Originally, it was thought that some actual examples from the Vietnam war and national strategic policy studies would be included. Nay had (has) particularly arresting examples that use an earlier variation of the approach. The examples are over ten years old. It turned out, however, that the examples were still heavily classified and could not be used. Thus, we were left with CM, instead of CIA or DOD. Apparently, those comparisons of rhetoric with reality from the war and from national strategy have a long hall-life and have still not decayed and cooled enough to be used.

As the reader can see at this point, this is not a book about budgeting, economic analysis, mathematical analysis, statistics, policy analysis, regulatory analysis, or even modeling and evaluation. And yet each and every step outlined seems to be a precursor to success in any of those fields. All require models of the structure of reality with measures of expectations firmly anchored to such models. The steps are also fairly easily mastered and applied, especially at the level required to examine the results of the lengthier, costlier, and more-detailed approaches discussed before.

At the time of this writing, the country is in the midst of reexamining policies and operations of government in energy, regulation, social programs, defense and international strategy, health care and health-care financing, and the national economy. All are complex systems subject to the approaches laid down here. We believe that each group needs to operate with at least one eye on reality.

We implied in the preface that the secrets of this book were applicable to most work in most government operations. Now that we have passed them on to you, make your own assessment and let us know the results.
Appendix

The Creation and Development of EA

The creation of EA began in the Program Evaluation Group at the Urban Institute during 1970-1973. The group, directed by Joseph Wholey and Deputy Director Garth Buchanan, had completed a survey of evaluation across the federal government and found some startling discrepancies between rhetoric and reality. Each new policy issue that developed seemed to create a battleground between the policy analysts and the people who did actual field work, measurement, and detailed examination of program operations.

Thus, the group set out to consider both the policy expectations and the program realities in a series of quick studies that encompassed manpower training programs; a national child-health demonstration program; the first eight grades of the Atlanta, Georgia, public schools; and the programs of the NIMH. The studies were to proceed to different levels of completion and detail, but all made an attempt to address both policy expectations and equivalency in the field and to compare the two for subsequent use by policymakers. All of these studies produced information and comparisons that were addressed by those in charge. Most helped to shape subsequent arguments and policies in their particular areas as well as arguments over when evaluation was appropriate and how it was to be usefully accomplished.

In the manpower work,[1] a summary of resulting views appeared in 1972,[2] and a synthesis of cost/benefits of these programs appeared in 1973. The cost/benefit synthesis was actually completed at the request of Senator Proxmire in June 1971, but enough controversy surrounded it so that it did not appear in public print until nearly eighteen months later.[3] Because several members of the group had run into similar controversy some years earlier after laying out the national strategic defense posture in terms of rhetoric and reality, the reactions came as no surprise.

The Atlanta school project resulted in a systemwide performance-signaling system that was tested over several years in the public schools.[4]

Each study report was shaped partially by the context in which it was done. The child-health study probably provided more EA-type interactions between high officials, program officers, and field-site findings than any of the others, especially during its first year and a half.[5] It made a full contrast between expectations and actual results.

The study at the NIMH produced some interesting contrasts between program rhetoric and actual practice and raised several issues about whether the rhetoric could actually be evaluated against the projects.[6]

The idea of making one cornerstone of the EA approach focus on whether rhetoric and reality were close enough to each other to be evaluable was first developed by the late Don Weidman after he joined the group. Joe Nay and John Scanlon coined the term evaluability assessment, but the term did not come into wide use immediately and the
Public Administration Review article (which was written earlier to meet the publication date) still finds us talking rather awkwardly about
preassessment of evaluability and skirting the idea that unevaluable programs often needed more management.[7] John Waller and John Scanlon drew together some of the implications for evaluation design in 1973.[8]

Tom White provided technical underpinnings to most of the studies performed and prepared many exceptional (and largely unpublished) bounding, clustering, and analytic critiques. In only two weeks, he turned the Phillips curve on its head and, in a couple of months, bounded the likely effects of manpower programs, results not highly popular at the time. Alfred Swartz brought reality to study after study by rushing out to examine reality in detail, by insisting that others do likewise, and by counseling sagely.

Pamela Horst and Richard Schmidt provided the fine writing and turns of phrases that made written ideas and concepts sing instead of plod, in addition to their regular work. Peter Bloch originally issued the challenge to create a set of research items that was readable by everyone.

Tom Kelley critiqued study after study with humorous responses that reminded us that people in the bureaucracy (where he was) could not do things quite as quickly or easily or safely as we could recommend them.

Richard Zamoff taught us how to use the telephone quickly, cheaply, but scientifically in order to gather rapid feedback information up and down the layers of complex organizations and programs.

Peg Kay had come to the problems of successfully mating expectations and operations in 1972 when she was an administrator in a large midwestern university. In 1973 she joined the staff of the Urban Institute’s Cable Television Information Center and in 1975 became principal investigator for a National Science Foundation study comparing seven research designs for the delivery of social services on cable TV.[9] During the course of that work, it became evident to her that in many cases social-science-evaluation methodologies were being applied to programs that were simply unevaluable. She compared notes with members of the Program Evaluation Group and early in 1976, left the cable center to work with the group, learning EA by coauthoring (with Joe Nay) the revisions of more than a dozen working papers that made explicit much of the early conceptual work on EA.

As the approach began to develop, thirty or more people inside and outside the group contributed to its development. John Wailer and Joe Wholey assisted a task force addressing LEAA’s block-grant program in 1973 to 1974, and the recommendation about finding out what was happening in this program led to the National Evaluation Program (NEP) and the Model Evaluation Program at the National Institute for Law Enforcement and Criminal Justice (NILECJ).[10] Wholey was the first director of the NEP project. Joe Nay stepped down as Wholey’s deputy in the Program Evaluation Group in 1974 to spend full time on the development of the NEP. Over the next several years, through the NEP, a
modified version of the basic approach was introduced by Richard Barnes of NILECJ to over 300 potential research grantees nationwide and a number of studies of LEAA’s field programs conducted by the grantees. This led to much dissemination, testing, and revision of ideas and approaches.[11]

A review by Knezo, at the Congressional Research Service, and various articles provided by group members provided exposure publicly, but the primary means of distribution was word of mouth, Xerox copies, work descriptions, reports, working papers, and examples.[12] A review by Tien, a book by Wholey, and a manual by Schmidt, Scanlon, and Bell have appeared more recently.[13]

The approach matured through short and long applications in a variety of federal departments, state, and local governments. Beginning in the mid1970s, practitioners of the methodology have spun off from the Program Evaluation Group to direct and participate in the management, planning, and evaluation of governmental operations, to form new organizations, and to join other, established organizations interested in extending the use of methodology. The Performance Development Institute (PD!) was formed in 1978 with the majority of its members from the Program Evaluation Group.

The Program Evaluation Group alone produced over a hundred reports, papers, and working papers. As applications began to spread in the mid-1970s, we have not been able to keep complete tabs on the large numbers of papers and books that drew upon the material or apply it. This book began in 1973 under Garth Buchanan, who left to become a vice-president in the Manville Corporation. Nay inherited the incipient book, along with over a dozen completely different and fairly incompatible outlines that had been developed by the group. Kay, bringing literacy and discipline to the effort as well as her own research contributions, joined the book effort in 1976. The chapters of the developing book became repositories for questions of approach and method that recurred again and again.

Joseph Wholey created a magic bubble within which the group was allowed to do this early work—that is, the people carrying out the work were protected by him, in large part, from the ideological and political attacks that sometimes result from displaying the rhetoric and reality of a program simultaneously. Wholey was able to maintain this protection until a critical mass of people began to build up who realized that serious problems existed in many governmental operations and in many of the theories behind them.

Only through Wholey’s special talent for this was an independent thinking group maintained during an ideologically turbulent time. The group grappled with the real operations of government, examined what was happening, what—according to theories and rhetoric—should have happened, and distinguished sharply between the two. In addition, the
group was often able to announce its results. Out of this work came the methods outlined in this book.

Evaluation

The genesis of this book is the twenty continuously revised working papers called research items that were originally created through 1973-1977. The purpose of these items was to serve as a repository of methodology and explanations of issues that recurred again and again in successful applications of program evaluation. This has largely determined the book’s present focus.

Each application of EA had a context of its own, developed some part of the complete picture more than others, and was not usually funded to include methods development. Thus, much of the writing and revising of this material was accomplished evenings and weekends, with the help of volunteers for typing and review.

One item—on evaluation design—has been omitted from this book, partially because this book is about oversight and EA, but mostly because, when we tried to translate the item into a chapter, we discovered that we were writing another complete book. The authors are fond of that item, and perhaps someday it will be expanded into usable form.

The work with the NEP at the NILECJ deserves special mention because it came at a pivotal time in the development of EA and allowed extensive testing of the concepts against the programs of a large nationally distributed block-grant program. It followed early versions of the precepts laid down in this book. It could not have operated or been successful without support from many levels. LEAA Administrators Donald Santorelli and Peter Velde and Deputy Administrator Charles Work set up special arrangements to meet EA requirements. Institute Director Gerald Caplan and Office of Research Programs Director Jeffrey Alprin steadfastly provided the funding, allowed the approach to be held constant until it could be tested and improved, took the heat when there were problems, and used the results as they came in. Richard Barnes, NEP director for the first three years, and his staff of Michael Mulkey and Jan Hulla, enlisted for the duration of the project and for long years fought the kind of wars necessary to develop, implement, and improve a large multi-project, multiarea effort involving over thirty topics and grantees. Sheila Lucille Graham was a key participant, first at NILECJ and then at the Program Evaluation Group. John Waller gave continual counsel. An account of the early part of this work is given in “The National Institute’s Information Machine.”[14]

About late 1975, evaluation applications of EA began to multiply. Wholey has documented some of the experiences in bringing EA to DHEW (now DHHS).[15] It is beyond the scope of this book to touch upon them all. Perhaps a book on the development of EA will
someday do justice to the subject.
Because all of our projects were team efforts involving many stakeholders in and outside of government, hundreds of people have contributed to the development of these concepts and to their applications in major ways.

**Government Oversight**

Prior to 1973, most interaction with the Congress over this body of work had to do either with evaluation or with the policy or content implications of particular studies, although the manpower cost/benefit study mentioned earlier was an attempt to synthesize all existing manpower and cost/benefit information for use in oversight in that area. Our insistence that evaluation must be usable, useful, and used—and that methods were available to accomplish this—led to more calls for informal consultation with congressional staff on overall questions of oversight during 1972-1974. Title VII of the Congressional Budget Act of 1974 authorized evaluation approaches more geared to oversight than to social-science evaluation as then practiced. This generated somewhat more interest in better oversight of federal operations, but the numbers of interested parties remained small. The budget process was being developed, but the question of how oversight information was to be obtained for the budget process did not draw much interest, remaining a battleground for a few specialists.

Senators Brock, Mondale, Cannon, Bellmun, Nunn, and Roth and their staffs were supportive. Harrison Fox (serving on a variety of Senate staff assignments) ran a series of seminars and produced a variety of directly useful materials. In October 1975, Senator Leahy requested the GAO to conduct a test and dry run of an oversight approach and submitted a Senate resolution based in part upon our earlier *Public Administration Review* article.[16] The GAO tests were carried out during 1975-1976, and the two authors of this book spent Christmas vacation of 1976[17] (at the request of Comptroller General Elmer Staats) working with Robert Kershaw of GAO on a precursor of the report that GAO was later to issue (much improved by GAO) in November 1977.[18] Testimony was also given in October 1977 before the Senate Committee on Human Resources to support the move for fuller oversight.[19]

The GAO continued to incorporate the EA approaches that had been developed, not only into their evaluation but also into their audit work over this time period. While little published, acknowledged, or quoted, the work of both the Program Analysis Division and of several audit teams in the operating divisions at GAO has broadened not only applications but also our theoretical reach. Michael Scriven of the Stanford Research Consortium said that the cost/benefit to researchers of several of these GAO publications (including Federal Program Evaluation: Status and Issues[25]) is infinite since the cost of obtaining them is nearly zero and they contain a number of very useful insights.[20]
The Canadian auditor general has also adopted the EA techniques and approaches.

Overall, at this writing, the oversight of government operations seems to be gathering steam, although much remains to be accomplished.

Regulation

In 1977, Nay, Waller, and Scanlon headed up a team that won a competition held by the ETIP at the Department of Commerce. This contract was for support in setting up what became the Regulatory Processes and Effects Project, a multiagency regulatory-reform effort. Regulatory-reform work through this project is still continuing at the Regulatory Center at PDI.

One of our continuing objectives is developing methods for the operation and change of large governmental operations. The regulatory work—over the years—is developing such approaches and examples. Because most existing social-science methodology is weak in just this area, four years of twin development (both adapting EA into regulation and assisting with actual experiments in regulatory change) is adding to the methods tool kit. The effort to effectively manage existing regulatory programs, to assess needed changes, to find proper initiatives for accomplishing the changes, and to assist in carrying them out has greatly broadened our understanding of the basic approach.[21]

It should be noted here that one of the modifications made has been to include additional levels (see chapters 9 and 14) for the functional modeling of the product regulated, its environment, the industry, and the direct regulator/regulatee boundary.

After four years of adapting EA to regulatory experiments, it has become even clearer how important it is to track causal flows and functions if causality is to be claimed. Attributions of cause and effect should be regarded with suspicion when the studies do not include the organizational settings or are done without a knowledge of the underlying frameworks. Development of any regulatory situation seems to require a framework for both regulatory expectations and for the measurement and analysis of regulatory processes and effects. This is the tedious and unsung labor of effective regulatory experiments.

Planning and Management

While Nay was extending EA into the regulatory arena, Kay was exploring its use as a tool for planning and control. She was particularly interested in simplifying its procedures so that it might be used by managers and field practitioners who were not necessarily conversant with either social-science methodologies or the language of EA itself.

Concurrent with her work with the Program Evaluation Group, she
became a fellow-in-residence at the Office of Telecommunications Policy/Executive Office of the President, where, in 1977, she wrote the
first EA-based planning manual.[22] That manual was directed toward planners with social-science backgrounds and a high degree of intellectual curiosity about all things methodological.

In 1978, Kay joined the National Rural Center as director of special projects and concentrated the efforts of her department on translating the methodology into a planning tool for managers and field practitioners. During that year Kay directed an unofficial consortium composed of the Department of Agriculture’s Rural Electrification Administration, the Heins Telephone Company, and the county government of Lee County, North Carolina. Kay developed a field manual for use in planning for rural telecommunications needs, and the other consortium members tested it in Lee County, under the field direction of Tom Wolfsohn.[23] No members of the field team (including Wolfsohn) had had previous experience with EA; the interviewers were telephone company employees.

Later that year, Kay began work with the Emergency Land Fund (ELF), a regional organization formed for the purpose of keeping Southern farmland in the hands of small farmers. Kay had been asked to develop a management system for ELF, and again she relied on a field team previously untrained in EA. By the end of the ELF project, it had become clear that, with some modifications, EA provided an excellent framework for planning and control systems.

In 1979, Kay and Joseph Brooks of ELF developed a workshop on EA as a planning and control tool, and the workshop, Program Planning and Management, was held in January 1980 under the joint auspices of the National Rural Center, ELF, and PD!. The four-day workshop was attended by federal- and local-government officials as well as members of private-sector service organizations (for example, a Long Island, New York, Catholic charities unit). This served to disseminate the methodology into some odd corners of the nation.

Late 1979 saw the planning/management version of EA take hold in the federal government. Duane Urban’s planning and control group at the Federal Insurance Administration (FIA) uses the methodology for that purpose as does Peg Kay’s group (of course) at the ICST at the National Bureau of Standards.

It should be noted that there are sharp differences between the way the process is used for a government agency’s internal-management purposes and the way a contracted EA is run as the basis for evaluation design. Foremost, the internal managers do not report to, negotiate with, and/or manipulate those in charge; they are those in charge. Usually that means that they are not social scientists, but practitioners, and as such that they tend to take shortcuts and eschew jargon. Neither Kay’s nor Urban’s group use words like testable model in the office. They implement the concepts in the office and talk about them at professional society meetings.

Another major difference is that internal managers are not protected by a work statement contained in a contract. They may intend to spend a
year or two drawing equivalency functional models. The road to serious budget cuts are paved with intentions like that, however, and they must respond when their congressional subcommittee or when OMB calls. What procedural niceties survive their negligence do not often survive their panic.

It should be evident by now that EA, used as a planning and control tool, is both methodology and—to use the words of FIA's Leo Bosner—a state of mind. Both organizations are continually matching the rhetoric of their plans with the reality of their operations and adjusting one or the other as appropriate. Both groups also report that this has worked well for them.

Notes


Appendix


21. A list of publications is available from PDI.


<table>
<thead>
<tr>
<th>Term</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADMHA. See Alcohol, Drugs, and Mental Health Administration</td>
<td></td>
</tr>
<tr>
<td>ADP. See Automatic data-processing centers</td>
<td></td>
</tr>
<tr>
<td>Accuracy and error, 65</td>
<td></td>
</tr>
<tr>
<td>Air-pollution-offset bank, 188—190, 200, 201—202, 204</td>
<td></td>
</tr>
<tr>
<td>Alcohol, Drugs, and Mental Health Administration, 81, 143</td>
<td></td>
</tr>
<tr>
<td>Alprin, Jeffrey, 228</td>
<td></td>
</tr>
<tr>
<td>Analytic integrity and political efficacy, 96</td>
<td></td>
</tr>
<tr>
<td>Anzelmo, Sam, 176</td>
<td></td>
</tr>
<tr>
<td>Ashby, Ross W., 36</td>
<td></td>
</tr>
<tr>
<td>Atlanta public-school project, 50—53, 225</td>
<td></td>
</tr>
<tr>
<td>Automatic data-processing centers, 37-38, 50, 168</td>
<td></td>
</tr>
<tr>
<td>Baltimore criminal-justice system, 176</td>
<td></td>
</tr>
<tr>
<td>Barnard, Chester, 20</td>
<td></td>
</tr>
<tr>
<td>Barnes, Richard, 227, 228</td>
<td></td>
</tr>
<tr>
<td>Behavior patterns, 109—110</td>
<td></td>
</tr>
<tr>
<td>Belief structure, 104—105; changing of, 167</td>
<td></td>
</tr>
<tr>
<td>Bell, James, 227</td>
<td></td>
</tr>
<tr>
<td>Bellmun, Senator, 229</td>
<td></td>
</tr>
<tr>
<td>Black-box models, 40—41, 50—51, 149</td>
<td></td>
</tr>
<tr>
<td>Black-hat audits, 179</td>
<td></td>
</tr>
<tr>
<td>Blau, Peter, 121 n.4</td>
<td></td>
</tr>
<tr>
<td>Bloch, Peter, 226</td>
<td></td>
</tr>
<tr>
<td>Bosner, Leo, 232</td>
<td></td>
</tr>
<tr>
<td>Bounding constraints, 7—8, 110—Ill</td>
<td></td>
</tr>
<tr>
<td>Brock, Senator, 229</td>
<td></td>
</tr>
<tr>
<td>Brooks, Joseph, 231</td>
<td></td>
</tr>
<tr>
<td>Buchanan, Garth, 225, 227</td>
<td></td>
</tr>
<tr>
<td>CAI. See Computer-assisted instruction</td>
<td></td>
</tr>
<tr>
<td>CHMC. See Community Mental Health Center Program</td>
<td></td>
</tr>
<tr>
<td>Cable television, 39, 226</td>
<td></td>
</tr>
<tr>
<td>Cannon, Senator, 229</td>
<td></td>
</tr>
<tr>
<td>Caplan, Gerald, 228</td>
<td></td>
</tr>
<tr>
<td>Cause/effect, 157 n.l</td>
<td></td>
</tr>
<tr>
<td>Characteristics to be measured, 61</td>
<td></td>
</tr>
<tr>
<td>Child-health project, 126—133, 185, 201, 202, 203, 225</td>
<td></td>
</tr>
<tr>
<td>Citzen participation, 104</td>
<td></td>
</tr>
<tr>
<td>Clean Air Act, 185—188</td>
<td></td>
</tr>
<tr>
<td>Community Mental Health Center Program, 81, 143</td>
<td></td>
</tr>
<tr>
<td>Comparison, quickset, 197—198</td>
<td></td>
</tr>
<tr>
<td>Comparison measures, 13, 20</td>
<td></td>
</tr>
<tr>
<td>Comparisons and reconciliations, 125; between families of models, 132—133, 155; within testable family, 129—132</td>
<td></td>
</tr>
<tr>
<td>Comparisons and tests of evaluability, 180—190</td>
<td></td>
</tr>
<tr>
<td>Comptroller General of the United States, 134—136, 175</td>
<td></td>
</tr>
<tr>
<td>Computer-assisted instruction, 144—156, 159—171, 180—184, 200, 201, 202, 203—204</td>
<td></td>
</tr>
<tr>
<td>Congressional Budget Act (1974), 229</td>
<td></td>
</tr>
<tr>
<td>Congressional Research Service, 227</td>
<td></td>
</tr>
<tr>
<td>Control measures, 13</td>
<td></td>
</tr>
<tr>
<td>Convergence, 168—170</td>
<td></td>
</tr>
<tr>
<td>Day-care programs, 39</td>
<td></td>
</tr>
<tr>
<td>Department of Energy Organization Act (1977), 138</td>
<td></td>
</tr>
<tr>
<td>Detail, level of, 167—168</td>
<td></td>
</tr>
<tr>
<td>Direct-intervention domain, 10-13; functional model of, 36—38, 53—55</td>
<td></td>
</tr>
<tr>
<td>Discussion: content/semantics/scope, 103—104, 154 Domains, 5, 8—15, 20 1—202</td>
<td></td>
</tr>
<tr>
<td>Drucker, Peter, 106, 107 n.4</td>
<td></td>
</tr>
<tr>
<td>E.A. See Evaluability assessment</td>
<td></td>
</tr>
<tr>
<td>ELF. See Emergency Land Fund</td>
<td></td>
</tr>
<tr>
<td>EPA. See Environmental Protection Agency</td>
<td></td>
</tr>
<tr>
<td>ETIP. See Experimental Technology Incentive Program</td>
<td></td>
</tr>
<tr>
<td>Effectiveness and efficiency, 106</td>
<td></td>
</tr>
<tr>
<td>Einstein, Albert, 69</td>
<td></td>
</tr>
<tr>
<td>Emergency Land Fund, 231</td>
<td></td>
</tr>
<tr>
<td>Enabling interventions, 13, 55</td>
<td></td>
</tr>
<tr>
<td>Environmental errors, 66—67</td>
<td></td>
</tr>
<tr>
<td>Environmental Protection Agency, 185, 188</td>
<td></td>
</tr>
<tr>
<td>Equal-interval scales, 62</td>
<td></td>
</tr>
</tbody>
</table>
Equivalency functional model 126—129, 159; preliminary, 161; revision of, 167
Equivalency information, 159—171; and testable information, 126—129
Equivalency models, 14, 125, 159, 221; building of, 160—166; comparisons and reconciliations with testable models, 132—133, 216, 219, 222; convergence with testable models, 168; and reality, 208—214
Equivalency summary statement, 129, 159, 161
Equivalency-to-reality functional models, 84—85
Error, kinds of, 66—67
Evaluability, comparisons and tests of, 180—190
Evaluability assessment, 123—141; creation and development of, 225–232 Evaluable, defined, 176
Evaluable model, 193; comparisons and tests of Evaluability, 180—190; and convergence, 168; and hybrid model, 170—171
Evaluable programs, 176—179
Evaluation, 228—229; defined, 3, 7; and purposeful behavior, 3—7, 19—29
Evaluation domain, 8—10
Evaluator: as perceived by those in charge, 87; role of, 16—18
Expectations, 17, 20; unrealistic, 136—137, 138
Experimental Technology Incentive Program (Commerce Dept.), 230
FIA. See Federal Insurance Administration
Federal Communications Commission, 39 Federal Evaluation Policy, 3
Federal Insurance Administration, 231
Feedback loops, 20—27; and equivalency models, 160; and governmental organization, 76
Flow diagrams. See Functional models
Flows, essential/unifying, 35, 43, 47, 50, 159, 166, 202, 210
Index

Information flows, 101—102
Input/output models. See Black-box models
Institute for Computer Sciences and Technology (National Bureau of Standards), 37, 50, 159—160, 166, 168, 231
Instrument errors, 66
Jefferson County, Louisiana. See Air-pollution-offset bank
Kay, Peg, 226, 227, 230—231
Kelly, Tom, 73, 226
Kershaw, Robert, 229
Kierkegaard, Soren, xiv
LEAA, 117, 226—227
Leahy, Senator, 229
Lee County, North Carolina, 231
Lessinger, Leon, 161—163
Lindsay, John, 137
Logic Models, 33, 36, 40-42, 51, 52-53, 75. See also Testable logic model
Louisville. See Air-pollution-offset bank
Management. See Those-in-charge domain
Manpower-training programs, 118, 120, 225, 226
Manville Corporation, 227
McNamara, Robert, 119, 137
Measure and scale selection, 61—63
Measurement, 45, 47, 49, 59—70, 202—204; steps in obtaining a 60—65
Measurement Instrument, 63—64
Measurement models, 33, 45, 49
Measurement points, 45, 47
Measurement precision, 64—65
Measures and comparisons, acceptability of, 91—96, 119—120
Michel son and Morley, 69 Michigan Office of Criminal Justice Programs, 59
Model-building process, 69—70 Model cities program, 218
Model Evaluation Program (NILECI), 226—227
Models, 35-55; comparisons and reconciliations between, 132—133, 155; convergence of, 168—170; families of 125
Mondale, Senator, 229
Money-generating market, 110-111
Mulkey, Michael, 228
Mystery model, 38-42
NILECI. See National Institute for Law Enforcement and Criminal Justice
NIMH, 117, 225
National Evaluation Program (NILECI), 226—227, 228
National Institute for Law Enforcement and Criminal Justice, 226—227
National Rural Center, 231
National Science Foundation, 226
Nay, Joe, 226, 227, 230
New York City, 137, 179
Newton, Isaac, 69
Nominal scales, 62
Nunn, Senator, 229
OPM. See Office of Personnel Management
OSHA. See Occupational Safety and Health Administration
Occupational Safety and Health Administration, 155
Office of Child Development, 39
Office of Education, 161, 184
Office of Personnel Management, 212—213
Office of Telecommunications Policy/Executive Office, 230
Operational errors, 67
Ordinal scale, 62
Organizational behavior. See Purposeful behavior
Organizational boundaries, 7—8, 110-111
Organizational hierarchies and discontinuities, 139—140
Organizational incentives, 109, 110—113; effects on the individual, 113—117
Organizational levels, 79—81
Outcome measures, 11, 45
Outriders, 154
PD. See Position description
PDI. See Performance Development Institute
Paperwork Reduction Act (1980), 203
Pennsylvania Electric Company, 185—188, 200, 201, 202, 204
Performance Development Institute, 185, 227, 230, 231
Performance measurement, 20
Planning and management, 76—78, 230—232
Plato, 69
Point of intervention: defined, 10; and organizational incentives, 112—113
Political efficacy and analytic integrity, 96
Position description, 212—213
Process measures, 11, 45
Program Evaluation Group (Urban Institute), 49, 50, 225, 226, 227
Program Planning and Management, 231
Proxmire, Senator, 225
Public Administration Review, 226, 229
Purposeful behavior, 109—20; and evaluation, 3—7, 19—29
Quickset comparison, 197—198
RANN. See Research Applied to National Needs
RIF. See Reduction in force
Randomization, 184
Ratio scale, 62—63
Reagan, Ronald, 138
Reality: and equivalency model, 208—214; and rhetoric, 225—226, 227
Reduction in force, 212—213
Regulation, 230
Regulatory Processes and Effects Project, 230
Regulatory reform, 230
Reinforcement, 119—120
Relativity theory, 69
Reporting forms, 63—64
Research Applied to National Needs (National Science Foundation), 112
Residual errors, 67
Review teams, 197
Rhetoric: and reality, 225—226, 227; and writing, 214
Rhetorical descriptions, 210—214
Rhetorical models, 75, 203
Roth, Senator, 229
Rural Electrification Administration, 231
Rural telecommunications project, 231
Santorelli, Donald, 228
Say’s Law, xx n.3
Scales, 6 2—63
Scanlon, John, 59, 226, 227, 230
Schmidt, Richard, 226, 227
Scoping decisions, 103—104, 154
Scriven, Michael, 229
Self-descriptions, 212—213
Self-descriptive models, 214—216
Semantics: of discussion, 104; multiple levels of, 81—85, 151—154
Sequential purchase of information, 133—134, 195—206
Simon, Herbert, 19, 109
Small farmers: land program, 154, 168
Staats, Elmer, 229
Stockman, David, xvii
Stogdil, Ralph, 70 n.4
Surrogate-intervenor models, 38
Swartz, Alfred, 226
Systemic errors, 66—67
Telecommunications high-school-equivalency project, 59—60
Testable descriptions, 210—214
Testable functional model, 75, 126, 143, 149—151; comparison and
reconciliation with testable logic model, 129—132; preliminary, 160; and testable information discussion and validation, 155
Testable information, 143—155; discussion and validation of, 151—155; and equivalency information, 126—129
Testable logic model, 126, 129, 143, 144—149; comparison of equivalency summary statement to, 161; comparisons and reconciliations with testable functional model, 129—132; rudimentary, 159; and testable information discussion and validation, 151—154
Testable measurement model, 75 Testable models, 14, 16, 17, 75—76, 125, 143, 221—222; comparisons and reconciliations with equivalency models, 132—133, 216, 219, 222; comparisons and reconciliations within family, 129—132; convergence with equivalency models, 168; emending of, 167; sources of information, 216—219
Thompson, James, 96 n.4
Those-in-charge domain, 13—15, 73—74; of government agencies, 75—85; what is acceptable to, 87—96
Tien, James M., 227
Time-allocation patterns, 105
Tuberculosis-testing project, 139, 169
Type III error, 59—60, 132, 136
Type IV error, 133, 137, 144
Type V error, 161
Uncertainty principle, 69—70 United States: Agriculture Department, 231; Commerce Department, 134, 230; Congress, 111—112, 118, 135—136, 175, 229; Energy Department, 138; Labor Department, 118
Universes, 200—201
Urban, Duane, 231
Urban Institute: Cable Television Information Center, 226; Program Evaluation Group, 49, 50, 225, 226, 227
Velde, Peter, 228
Vietnam War, 119—120, 137, 169, 179, 224
Virginia community mental-health program, 82—83
Waller, John, 226-227, 228, 230
Washington subway-fare system, 63—64
Weatherization program, xviii—xix
Weidman, Don, 226
White, Tom, 226
Wholey, Joseph, 3, 225, 226—227, 228
Wildavsky, Aaron, 87, 88, 96
Wilson, Charles, 119
Wolfsohn, Tom, 231
Work, Charles, 228
Written information, 214